

Round objects

Planets are spherical, and the International Astronomical Union's attempt to make this part of their definition has merit.

There was once a prissy British civil servant who, when he came across a passage in a memo that displeased him, wrote "round objects" in the margin as a synonym for something ruder. This arch circumlocution was lost on the bluff minister he served, who fired back a query as to who this Round fellow was, and why he objected so much.

We can expect there to be plenty of members of the International Astronomical Union (IAU) who, reading the proposed new definition of a planet offered to them by their executive committee, will want to scrawl something equally rude and rather blunter in the margin — and will want to make their objections heard, possibly quite vociferously, at their general assembly in Prague this week (see page 724).

We understand and, to some extent, sympathize. But we would suggest that, instead, they acquiesce in the new definition, which will have the effect of increasing the number of planets in the Solar System to 12, and open the doors to more. They should do this for two reasons: it is not a bad definition; and it will at least stop the rumbling debate over the status of Pluto.

In the 1990s, it became clear that Pluto, the most newly discovered planet, was the most conspicuous of a crowd of icy 'trans-neptunian objects' (TNOs), some of which might well be larger. There was an obvious historical parallel to this situation with asteroids in the nineteenth century. When it was found that there were dozens of asteroids, Ceres, the largest and first discovered, was demoted from its position as a proper planet; it is now a 'minor planet' along with all the other asteroids. Pluto, it was argued by analogy, should be a minor planet with the rest of the TNOs on similar grounds.

This proposal sparked a degree of public debate that irritated many astronomers, who felt that the question of whether a particular body gets called a planet or not is of no scientific interest whatsoever. Still, the IAU decided that it should try and resolve the matter: planets

loom large in the public imagination, and it seemed only reasonable for astronomers to be able to say whether a new discovery (or for that matter an old friend) was a planet or not.

The IAU's proposal is that the term 'planet' should apply to an object that has a sufficiently strong gravitational field to have pulled itself into a spherical shape, that is in orbit around a star, but that is not a star itself. This lets in Pluto and 2003 UB₃₁₃, a TNO that is a touch bigger and not yet equipped with an IAU-approved name. It also readmits Ceres. And in the most peculiar aspect of the whole business, Charon, previously considered to be a moon of Pluto, will become a planet in its own right. Moons, however spherical, will remain satellites, not planets, in the IAU's eyes. But because the centre of mass of the Pluto–Charon system lies outside the body of Pluto, Charon, although tiny compared with, say, Neptune's moon Triton, qualifies as a planet.

Nine more TNOs, and three more asteroids, will become candidate planets, pending further investigation of how spherical they are. More planetary TNOs may follow, when discovered. To tidy things up, the minor planets will get renamed: those that don't have enough of a gravitational grip on themselves to be proper planets will now be 'small Solar System bodies'.

All this will doubtless lead to ructions. But it is at least a coherent approach, and it has a fairly clear basis in physical properties. It has been convenient to have a small and easily memorized number of planets in the Solar System, but convenience is not the only thing that counts. The effects of mass define (unofficially) the upper limits of the planetary realm; anything big enough for fusion is a star. It is fitting, then, that mass should define the lower limit too. This, we think, adds up to a case for IAU members to accept the proposal. ■

"All this will doubtless lead to ructions, but it is at least a coherent approach, and it has a fairly clear basis in physical properties."

Revival in Iran

Whatever its motivation, Iran's support for education and science is to be welcomed.

In eleventh-century Persia, it is said that three school friends pledged to serve their country and share their fortunes. Very different fortunes, it turned out.

Nizam al-Mulk became prime minister to two consecutive Persian kings. He built a network of roads across the country, and established the chain of 'Nizamiyya' schools, which taught theology, science and mathematics, adhering to a national curriculum.

Hassan-i Sabbah became the head of a fanatical religious group, the

Hashshashin, which operated an almost independent government, protected by a string of castles. The many attempts by Persian kings to overthrow the Hashshashin failed, and Nizam al-Mulk was eventually assassinated by Sabbah's followers.

Omar Khayyam became the greatest astronomer and mathematician of his age. He invented, for example, the Khayyam triangle — better known as the Pascal triangle, after Blaise Pascal who described it hundreds of years later. Khayyam also provided his country with a solar calendar, more accurate than the gregorian calendar we use today. And he became one of Persia's most popular poets.

In the millennium since the three school friends parted company, the country we now know as Iran has witnessed a sometimes glorious, often sad, political history. Along with the rest of the Middle East, Iran's scientific power declined as Europe's ascended with the



Renaissance. But the nation's cultural respect for study never died.

Science regained its foothold during the 1970s, under the Shah, even though his oppressive regime drove many intellectuals into exile. It faltered at the start of the Islamic revolution in 1979, but gained momentum in the 1990s when Iran became the most scientifically productive country in the Middle East apart from Israel. About 4,000 papers from Iran were published in 2005, according to the Institute for Scientific Information, compared with just over 500 in 1995. (*Nature's* first all-Iranian research paper was published last week.)

Perhaps the rise of science relates to the importance that Iran's government attaches to the development of nuclear technology. Many regard Iran's interest in these technologies with extreme suspicion. Nonetheless, Iran's embrace of science should be welcomed.

The educated young in Iran will still go their own individual ways, usually for good, sometimes for bad. But there is once again the opportunity for a privileged few to shine as scientists, if they can cope with the low pay and poor infrastructure that prevail outside the handful of elite institutions, and can sidestep the many problems caused by US sanctions.

One practical advantage for science in Muslim countries is the lack of direct interference of religious doctrine, such as exists in many Christian countries. There has never, for example, been a debate about darwinian evolution, and human embryonic stem-cell research

is constrained by humanistic rather than religious ethics. The Royan Institute in Iran was the first in the Middle East to develop a human embryonic stem-cell line, using spare embryos from its *in vitro* fertilization programme.

The recent dramatic rise in scientific productivity coincided with the relaxing of a stern Islamic regime under reformist president Mohammad Khatami. When hard-line Islamist Mahmoud Ahmadinejad was elected president last year, some scientists felt nervous, especially, no doubt, when he replaced the presidents of all the universities with worryingly inexperienced people. But the regime has so far shown a strong commitment to higher education. One of its first acts was to wipe out the debts accrued by universities, where female students now outnumber males, even in some areas of hard science. Ahmadinejad has also taken significant steps to prepare for an expansion of university student numbers. And he has not made cuts to research funds, which had increased over the past decade.

But will he maintain growth, given other pressing priorities in today's Iran? If not, many young scientists trained in the recent good years and now undertaking postdoctoral research abroad will have no prospects if they return home. An opportunity would be lost. So here's hoping that he avoids the need for these lines of Khayyam as a lament:

*Alas that spring should vanish with the rose
That youth's sweet-scented manuscript should close!* ■

Preventing cancer

More support is required to tackle obesity as a means of cancer prevention.

Twenty-five years ago, a landmark study by Richard Doll and Richard Peto concluded that 75–80% of cancers diagnosed in the United States in 1970 might theoretically have been prevented by altering environmental factors such as smoking, alcohol consumption and diet (R. Doll & R. Peto *J. Natl Cancer Inst.* **66**, 1191–1308; 1981). More recent work has added obesity and physical inactivity to the list of factors that increase cancer risk. Today, tobacco use accounts for 30% of cancers in the United States, obesity accounts for 15% and poor diet for up to 25%. Clearly, while avidly pursuing promising avenues of therapy (see page 735), the biomedical community and policy-makers need to tackle cancer prevention with just as much zeal.

The success of campaigns to combat melanoma and lung cancer are testaments to the impact that prevention can make. For example, the Australian government's exhortation to "slip on a shirt, slop on a sunscreen, slap on a hat" has led to melanoma having less of a health impact in Australia than in cloudier countries such as Britain. The incidence of smoking-related cancers has dropped sharply since the 1990s — the delayed effects of anti-smoking campaigns started in the 1960s and 1970s.

But much still needs to be done. Tobacco remains a problem, especially in the developing world, where consumption is climbing. The next biggest threat facing the developed world is the growing

epidemic of obesity. In the United States and Europe, for example, around two-thirds of the population is either overweight or obese. Research into how people form dietary and exercise habits will help to inform intervention campaigns, and research on nutrition will help to determine why a vegetable-rich, low-saturated-fat diet seems to provide protection against cancer. Studies on genetic susceptibility could one day help us all to tailor our lifestyles to suit our risk profiles.

But such research is likely to have little effect unless policy-makers are prepared to go further than just educating the public. Avoiding the Sun is straightforward and cheap. The same cannot be said of factors that affect obesity. Once associated with wealth and excess, obesity now disproportionately affects the poorer sections of society, because high-calorie, low-nutrient, processed food is often much more easily accessible than healthy alternatives. Governments need to help ensure that eating the recommended five portions of fruit and vegetables a day is a realistic aim for everyone.

Physical activity needs to be tackled too. Encouraging children and adults to take up sport is a start, but policies that encourage people to build exercise into their daily lives — town planning that makes walking to the shops or cycling to work easier and safer, for example — are likely to have a longer-lasting impact.

Our modern lifestyles are increasingly at odds with the environment in which our physiology originally evolved. Finding a truce between the two would go a long way to reducing the burden of diseases, such as cancer, that result. ■

"The success of campaigns to combat melanoma and lung cancer are testaments to the impact that prevention can make."

RESEARCH HIGHLIGHTS

MICROBIOLOGY

Gut response to toxin

Science **313**, 848–851 (2006)

Certain strains of the bacterium *Escherichia coli* manufacture a toxin that introduces double-strand breaks into mammalian DNA, according to new research. The damaged cells swell up and stop dividing.

Eric Oswald of the National Veterinary College of Toulouse, France, and his colleagues identified a gene cluster responsible for producing the toxin in different strains of *E. coli*. The cluster was found both in strains that cause infection and in strains that live in the human gut, including one used for intestinal 'probiotic' treatment.

The bugs may benefit from blocking intestinal cell division, as this could help them to colonize the gut. The DNA damage caused by the bacteria might predispose the host organism to cancer.

HUMAN GENETICS

Deletion diagnosis

Nature Genet. doi:10.1038/ng1853; 10.1038/ng1858 and 10.1038/ng1862 (2006)

A new clinical syndrome has been identified in individuals with learning difficulties. The syndrome, as yet unnamed, is caused by a very small deletion in chromosome 17. It is expected to explain about 1% of the undiagnosed cases of mental retardation.

The deletion was identified by three independent research teams using genomic tools. Bigger chromosomal abnormalities that cause mental retardation, such as Down's syndrome, can be seen under the microscope.

The deleted region is flanked by DNA sequences that are typical of other areas prone to deletion. It carries only about half a dozen genes. One of these is the so-called tau gene, *MAPT*, which is mutated in some neurodegenerative disorders, including Alzheimer's disease.

ASTRONOMY

Line of sight

Astrophys. J. preprint available at www.arxiv.org/abs/astro-ph/0606279 (2006)

Astronomers have used the circling dance of two stars to determine the distance to one of the Milky Way's neighbours, galaxy M33. The measurement pushes to its limit one of the few direct techniques for mapping space.

Alceste Bonanos of the Carnegie Institution of Washington and her colleagues studied how the brightness of the stellar pair changed

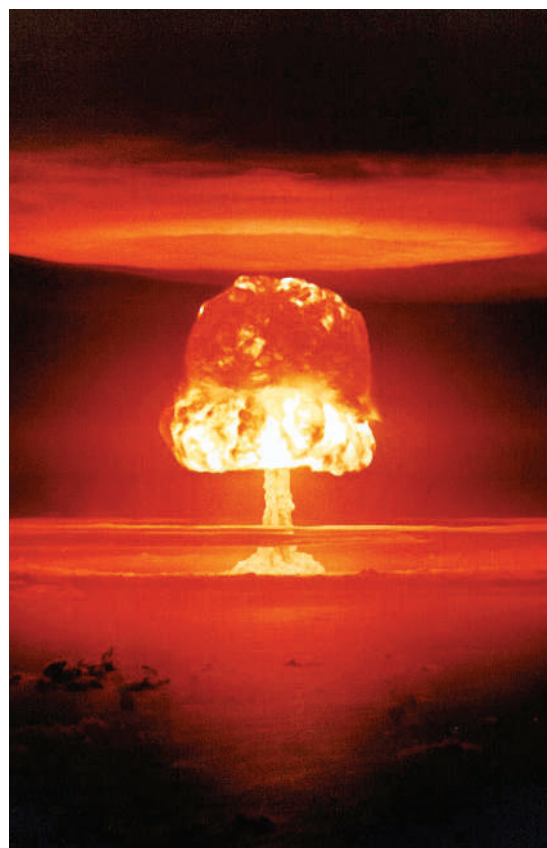
After the bomb

Proc. Natl Acad. Sci. USA **103**, 12564–12568 (2006)

A legacy of nuclear-bomb tests is helping to settle controversy about whether new neurons are born in the brain's neocortex throughout our lives.

The level of carbon-14 in the atmosphere (and thus the human body) was more or less doubled by Cold War bomb tests, then dropped after the 1963 Test Ban Treaty. A team led by Peter Eriksson of Gothenburg University and Jonas Frisén of the Karolinska Institute, Stockholm, Sweden, measured carbon-14 in autopsy tissue of people born before and after the bomb tests. Levels of the isotope in neocortex neurons corresponded to atmospheric levels at the time of the person's birth, suggesting that almost all neurons are born at that time. This supports Frisén's earlier work.

Furthermore, new neurons did not appear to be generated in the neocortex of cancer patients who took a compound that marks dividing cells.



as one star eclipsed the other. From the orbit's properties, the team estimated the two stars' masses, and thus their absolute brightness.

Then, by comparing the observed brightness to that predicted, they calculated the stars' host galaxy to be about 3.1 million light years away.

The result is at the upper end of the range obtained by other methods. It is also the greatest distance over which the binary technique has been applied: the stars would be unresolvable if any farther away.



CANCER BIOLOGY

Dog cancer has wolf roots

Cell doi:10.1016/j.cell.2006.05.051 (2006)

An unusual, contagious cancer that affects dogs probably emerged in wolves or their relatives between 200 and 2,500 years ago, according to researchers in the United Kingdom and the United States.

Canine transmissible venereal tumour (CTVT) is passed from pooch to pooch by transmission of the tumour cells themselves, usually when the animals have sex. Robin Weiss of University College London, UK, and his colleagues compared the genotype of the tumour to that of more than 400 dogs in 85 breeds. The tumour was most closely related to dogs such as wolves and huskies (pictured left). The researchers estimated the age of the cancerous cell line by counting mutations.

IMMUNOLOGY

Weakened zinc link

Nature Immunol. doi:10.1038/nri1373 (2006)

Zinc deficiency is known to impair the immune system, but the mechanism for this has been unclear. New work hints that zinc may have a signalling role in the development

of some immune cells.

Toshio Hirano of the RIKEN Research Center for Allergy and Immunology in Kanagawa, Japan, and his colleagues studied dendritic cells, which prime other immune cells to respond to a pathogen's proteins. They showed that the zinc concentration in dendritic cells fell when they were stimulated in a way that mimics infection. Moreover, artificially reducing the zinc concentration in dendritic cells stimulated them to mature as they do when faced with a pathogen. The researchers suggest that zinc deficiency may alter immune function in part by knocking this system off balance.

ANIMAL BEHAVIOUR

Fish farmers

Biol. Lett. doi:10.1098/rsbl.2006.0528 (2006)
Damsel fish, residents of coral reefs, are known to be conscientious farmers, cultivating crops of algae for their own consumption. A detailed study of the species' habitats has now revealed that some algae are equally dependent on these fish.

Hiroki Hata and Makoto Kato of Kyoto University, Japan, collected hundreds of samples of algae from both within and outside damselfish territories. One *Polysiphonia* algal species dominated the farms of the damselfish *Stegastes nigricans* and was found nowhere else.

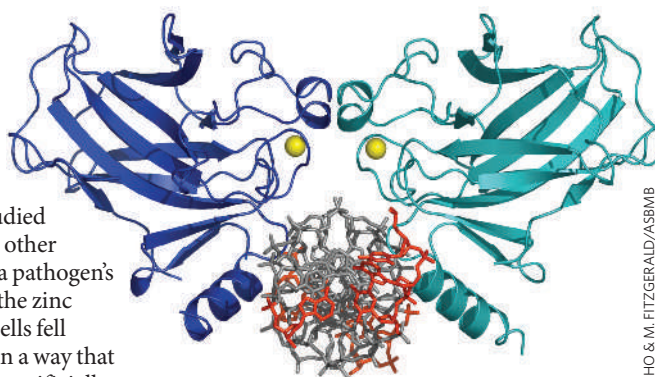
This suggests that the two species rely on each other: a situation described as obligate cultivation mutualism. Fungi-cultivating insects and crop-growing humans have a similar set-up.

CHEMICAL BIOLOGY

Improving nature

J. Am. Chem. Soc. doi:10.1021/ja061099y (2006)
As catalysts, enzymes are hard to improve on. But Ryan Mehl and his colleagues at the Franklin and Marshall College in Lancaster, Pennsylvania, have succeeded. By inserting a non-natural amino acid into the enzyme nitroreductase, they increased its activity.

Nitroreductase activates the anticancer prodrug CB1954, converting it to a toxic form. The researchers measured the effect of replacing a key amino acid in the enzyme's active site with one of eight different non-natural amino acids, each of which was predicted to improve binding of the prodrug. One of these made the enzyme 30 times more efficient than natural nitroreductase.



CELL BIOLOGY

Mutated in the middle

J. Biol. Chem. **281**, 20494–20502 (2006)

The so-called 'guardian of the genome', p53, protects against many cancers — unless mutations prevent it binding to DNA. Crystal structures of the p53–DNA complex show how this can happen.

Crystal structures of a single p53 protein bound to DNA have previously been used to explain the effect of some mutations implicated in cancer. But p53 actually binds in fours, as pairs of dimers. Ronen Marmorstein and his colleagues from the University of Pennsylvania in Philadelphia studied the structure of a p53 dimer bound to DNA (pictured above; DNA shown in red and grey).

They found that many faulty p53 proteins have mutations in the area where the two units of the dimer join. A model of the full p53 tetramer–DNA complex suggests that contacts between two dimers are less frequently mutated.

OPTICS

The mirror vanishes

Phys. Rev. Lett. **97**, 053902 (2006)

If you shine light on a silver mirror, you'd expect it to bounce back. But Ian Hooper and his colleagues at the University of Exeter, UK, have shown that, in the right circumstances, light can pass straight through.

They sandwiched a 40-nanometre-thick silver film between two glass prisms coated with thin films of zinc sulphide. In this arrangement, the refractive indices at the material interfaces are well matched, rendering the silver film partially invisible to a light beam — in the same way that a drinking glass may vanish when immersed in water.

Transmission peaked at about 35% for visible wavelengths — not perfect, because the zinc sulphide absorbs some light. The effect could be useful for making metal electrodes transparent in light-emitting diodes.

JOURNAL CLUB

Laurence Eaves

University of Nottingham, UK

A semiconductor physicist sees graphene transistors as a triumph for small-scale research.

Reading about a major scientific breakthrough from a small university group always gives me particular pleasure. Funding agencies tend to favour big headline-grabbing initiatives, but many research successes in condensed-matter physics, such as the development of magnetic resonance imaging and polymer light-emitting diodes, have come from small groups funded initially by modest grants.

Recent work from Manchester University, UK, on the electronic properties of graphene further demonstrates what a small, dedicated team can achieve. In 2004, researchers there fabricated a transistor from a few atomic layers of graphite (K. S. Novoselov *et al. Science* **306**, 666–669; 2004), a cheap and widely available material.

Why, I wondered, hadn't somebody thought about doing this before? And why were the electronic properties so good?

With the benefit of hindsight, one can see why it works. The Manchester team had the skill and patience to cleave graphite right down to a single atomic layer. An atomic sheet of graphite (graphene) contains few free electrons, so it is easy to control the current flow. Also, the electrons in graphene have a curious property: they move as though they have almost no inertial mass, so the device can act as a fast switch.

Recently, the same group showed that small ripples in graphene, like rucks in a carpet, scatter electrons via a subtle quantum effect (S. V. Morozov *et al. Phys. Rev. Lett.* **97**, 016801–4; 2006). By improving their fabrication technique, the team succeeded in removing the ripples, enhancing further the device characteristics.

Longer-term, we should see new applications of graphene transistors in electronics. Funding agencies and policy-makers, please note!

NEWS

Planets are round. Will that do?

Next week, by a simple show of hands at an astronomy meeting, Earth could go from being one of nine planets to one of twelve — with unknown numbers yet to be discovered.

A seven-member panel appointed by the International Astronomical Union (IAU) has recommended a new definition of a planet: any body in orbit around a star that is not a star itself nor in orbit around a much larger planet, and that is massive enough for gravity to have squished it into an approximately spherical shape.

The IAU planned to introduce the notion in a draft resolution on 16 August, after *Nature* went to press, and will discuss it on 22 August during an open session at its general assembly in Prague, Czech Republic. Two days later, if a simple majority of IAU members vote for it or a slightly revised version, that will become the official scientific word on the topic.

Under the definition, most objects with a mass greater than 5×10^{20} kilograms, and typically with a diameter greater than 800 kilometres, would qualify as planets. That would instantly elevate the asteroid Ceres and Pluto's moon Charon to planet status, as Ceres is roughly spherical and Pluto and Charon can

be described as a binary planet system. The Solar System's twelfth planet would be object 2003 UB₃₁₃, nicknamed Xena, one of many 'trans-neptunian objects' lurking on the Solar System's fringes.

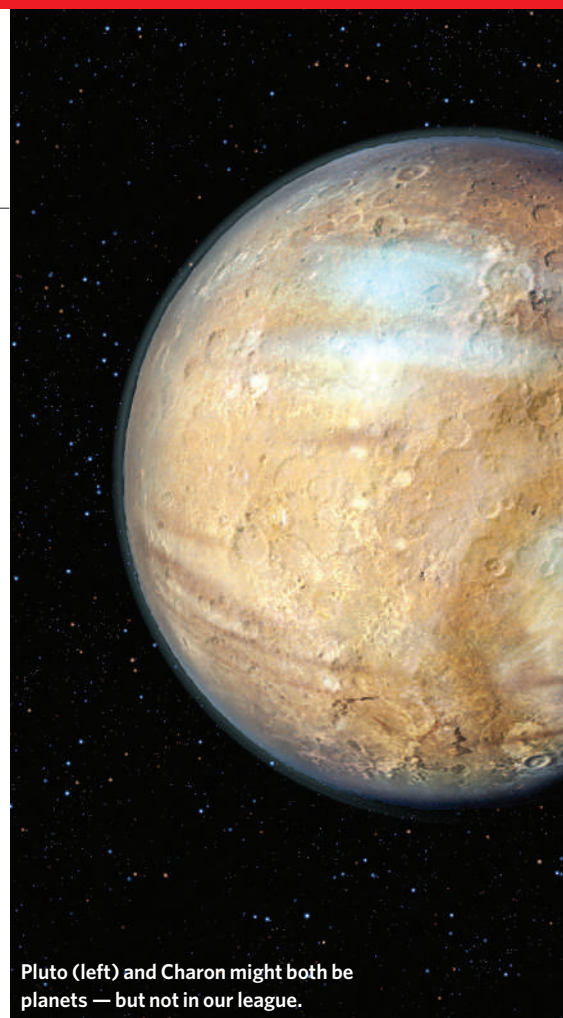
Not all are happy with the 'roundness' criterion, however. "It's not what I think of when I think of a planet," says Michael Brown, an astronomer at the California Institute of Technology in Pasadena and co-discoverer of UB₃₁₃. "I don't think of things that are a thousand times smaller than Earth."

The discovery of UB₃₁₃ beyond Pluto's orbit last year brought matters to a head. Astronomers have estimated that it is at least 2,400 kilometres in diameter — larger than Pluto, but smaller than any of the other eight planets.

All this demands a new definition of 'planet', some say. "If there is nothing accepted, we have to sit in the silly place where we are now: Pluto is a planet and UB₃₁₃ is not," says Iwan Williams, an astronomer at Queen Mary, University of London.

Williams chaired an earlier IAU committee that was meant to define a planet, but it deadlocked last November. "People came in with well-defined ideas of what a planet should be, expressed them strongly, and hardly anybody

"If there is nothing accepted, we have to sit in the silly place where we are now."



Pluto (left) and Charon might both be planets — but not in our league.

changed their mind," says committee member Alan Boss, an astronomer at the Carnegie Institute of Washington. Opinions were divided between three options to define a planet orbiting a star: any object with a diameter greater than 2,000 kilometres; any object massive enough for gravity to make it round; or any object that dominates its region of space.

Faced with that stalemate, the IAU appointed

D. VAN RAVENSWAAY/SPL

AIDS meeting urged to rethink prevention strategy

TORONTO

Scientists, governments and funding agencies need to revise their strategy for deploying and testing HIV-prevention methods, delegates to the XVI International AIDS Conference in Toronto, Canada, heard this week.

Although conventional prevention methods, such as the use of condoms, have had some success, advocates say that women need to have more control over prevention. Various new approaches, including the externally applied gels known as microbicides or drugs taken orally, show promise, said leaders in the fight.

Challenges remain, however, in bringing such methods to the clinic. "There are serious obstacles that could significantly delay, or even derail, critical prevention trials — including inadequate resources and capacity to launch and complete trials, and emerging ethical concerns," warned a group of scientists, doctors and activists called the Global HIV Prevention Working Group.

A 15 August report from the coalition called on governments and scientists to focus on the most promising interventions, set up trial sites, and coordinate their plans to avoid wasting time and money

duplicating efforts. It also advised the creation of a panel of experts that could provide ethical guidance for the trials. And it urged societies to prepare for the roll-out of new prevention methods if they prove successful.

The warnings aim to reduce the problems already hampering efforts to find alternative prevention methods. For example, activists' concerns, such as over whether people who contracted HIV during the trial would receive treatment afterwards, shut down two trials on preventive drugs in Cambodia and Cameroon in 2004 and 2005.

Scientists hope that interest

from big funding groups, such as the Bill and Melinda Gates Foundation, based in Seattle, Washington, could focus attention on these preventive efforts.

Trials are currently testing the effects on HIV infection of microbicides, orally-ingested drugs, male circumcision, barriers such as diaphragms and the treatment of herpes infections. But the studies are costly and difficult to run.

"We need to make strategic choices, and I hope Gates is going to give leadership in the way that is finally happening in the vaccine field," says Joep Lange, chairman of the International AIDS Society's



another planet definition committee, chaired by astronomer and historian Owen Gingerich of Harvard University. The group met on 30 June for two days at the Paris Observatory. "We converged relatively quickly after the first day," says panel member Richard Binzel of the Massachusetts Institute of Technology.

A definition based on gravity, the group concluded, made the most scientific sense. "It

involves the most physics," Binzel says.

To acknowledge that trans-neptunian objects are different to the other planets, the committee proposed a subcategory of planets known as 'plutons'. These are planets that take more than 200 years to orbit the Sun and would include Pluto, Charon and UB₃₁₃.

The number of planets is likely to rise as more large trans-neptunian objects are found. After Ceres, other asteroids may also be eligible. Such changes may upset generations who grew up with nine planets. "If I have any concern, it's that the public will accept this," says astronomer Ron Ekers, president of the IAU.

Binzel thinks there are ways around that. "My expectation is that children will memorize the eight classical planets, they will know that Ceres is a planet in the asteroid belt, and that there's a whole collection of planets out beyond Neptune, of which Pluto is the first."

Convincing fellow astronomers may not be easy either. "There will be a long line of people waiting for the microphone to denounce it," Boss predicted before he had heard details of the proposal. "I think there's a good chance nothing will be decided formally."

That prospect dismays those who have been working on the new definition. "This is a very good compromise," says Binzel, "and it's time to move forward."

Jenny Hogan

With additional reporting by Geoff Brumfiel

Discuss the IAU proposal on what makes a planet — and propose your own mnemonic for what could be the new 12 planets — on *Nature's* newsblog, blogs.nature.com/news/blog. See also the Editorial on page 719.



INTERNATIONAL AIDS MEETING

Read our conference report on the newsblog.

<http://blogs.nature.com/news>

G. ROBINS/AFP/GETTY

Homing in on the genes for humanity

Researchers have identified a gene that has changed rapidly during human evolution — a discovery that could be a step towards understanding what sets us apart from other animals.

The work, reported online this week in *Nature*, is some of the first to come from comparing the human genome with that of the chimpanzee, which was published last September. The gene, called *HAR1F*, does not directly code for a protein. Instead, it lies in a non-coding segment of the genome and produces RNA. Many geneticists believe that rapidly evolving non-coding regions harbour the secret of what makes humans different from our nearest primate relatives.

HAR1F is produced by cells in the brain called Cajal–Retzius cells, which regulate how the six layers of the cortex are laid down during development. The gene may interact with a protein called reelin, which plays a vital role in this layering. "But it's wild speculation," says geneticist David Haussler of the University of California, Santa Cruz, who led the study.

The group identified the gene by comparing the human genome with those of the chimpanzee, mouse and rat. Forty-nine segments, dubbed 'human accelerated regions' or HARs, showed sequence changes in the human version but not in the other animals. The greatest change was found in *HAR1*, which in humans had undergone 18 substitutions in comparison with the other animals when one or none was expected (D. Haussler *et al. Nature* doi:10.1038/nature05113; 2006).

What the gene does is a mystery, but there are some guesses. "Given that it's changed so dramatically only for humans, it might be involved in human-specific brain wiring," says Gerton Lunter at the University of Oxford, UK.

One thing is becoming clear: protein-coding genes may not be the movers and shakers of human evolution scientists once thought. "We should stop looking at proteins and start looking at non-coding DNA," says Lunter. "Everything points in that direction."

Kerri Smith

J. UZON/AFP/GETTY



This year's AIDS conference focused on new preventive methods.

Industry Liaison Forum.

Trial results can also be complicated when counselling, provided with the trial, cuts the rate of new infections. For instance, the first completed

trial of oral prevention drugs, reported at the meeting, tested whether the oral drug tenofovir prevented new infections in 936 women mainly in Ghana. But too few people became infected, so

it was difficult to tell whether the drug actually worked, said Leigh Peterson of Family Health International, which conducted the trial with money from the Gates Foundation.

The foundation has already led efforts to coordinate research on a microbicide, and helped write a strategy for microbicide development, due to be released on 17 August. Renee Ridzon, a senior programme officer at the foundation, said it is keenly aware of the need to coordinate prevention trials.

"The field needs to think hard about this," she said. "It's definitely on our radar screen." ■
Erika Check

ON THE RECORD

“People may have thought: ‘We have tapes of the Moon walk, we don’t need these.’”

Australian scientist John Sarkissian hunts for the original high-quality magnetic tapes, apparently misplaced, of the first Apollo 11 Moon walk.

“The gene won’t automatically make you a criminal.”

New Zealand researcher Rod Lea triggers controversy by reporting a high incidence among Maori people of a gene linked to aggression.



WHILE YOU'RE THERE

Those attending the International Astronomical Union's general assembly in Prague, Czech Republic, from 14 to 25 August should be sure to check out the astronomical clock, or Orloj, in the town square. The clock has been operating since 1410, on and off, and so the meeting's attendees may notice it is slightly out of date. Research done since its construction suggests that Earth orbits the Sun, not the other way around.

DATAPOINT

Turning the corner?

After years of being thwarted by immigration difficulties, international students may finally be returning to US graduate schools in greater numbers. A study in admissions found increases since 2005:

Life sciences	+1%
Physical sciences	+5%
Engineering	+26%

Sources: Sydney Morning Herald, Australian Associated Press, Council of Graduate Schools

Guilty, but no jail sentence for Russian scientist

Russian physicist Oskar Kaibyshev was given a six-year suspended prison sentence last week for exporting technologies with possible military use to South Korea. Human-rights advocates say the accusation is baseless and part of a series of prosecutions unjustly targeting Russian scientists.

Kaibyshev, the suspended director of the Institute for Metals Superplasticity Problems in Ufa, was also fined about US\$130,000 and banned from resuming his directorship for three years. Reports say he intends to appeal against the verdict.

In 2002, Kaibyshev's institute sent samples of aluminium alloys and a titanium product to a tyre company in Seoul, South Korea. The materials allow better high-pressure tyres to be made, but the prosecution said they could also be used in manufacturing missiles.

Agents of the Russian Federal Security Service first arrested Kaibyshev in March 2003 after seizing documents from a South Korean busi-

ness delegation, including representatives from the tyre company and the Korea Aerospace Research Institute, at the airport in Ufa. He was arrested again in February 2005.

Kaibyshev's supporters say the information he provided was public knowledge. "Every country has the right to defend its state secrets, but Kaibyshev's defence has clearly proved that all information he passed on was previously published in books and journals," says Eugene Chudnovsky, a condensed-matter physicist at the City University of New York who chairs the human-rights division of the New York Academy of Sciences.

Over the past decade, Russian courts have given lengthy prison sentences to a number of scientists convicted of spying. Prominent examples include physicist Valentin Danilov and military analyst Igor Sutyagin, who were sentenced to 14 and 15 years, respectively, in 2004. Several other Russian scientists are also being held in custody for alleged espionage and

The outlook for Amazonia is dry

Drought hit the Amazon last year, with devastating results. Rivers fell, fish rotted and routes to schools and hospitals were cut off. Now studies of the dry spell suggest that such conditions could become increasingly common.

The drought first caught scientists' interest because its cause was unusual. Dry spells in the Amazon usually occur in El Niño years, when warm water off the Pacific coast of South America sets up a pattern of circulating air that inhibits rainfall in the Amazon. But last year was not an El Niño year.

Instead, the drought was caused by a circulation pattern powered by warm seas in the Atlantic — the same phenomenon responsible for last year's unusually intense Atlantic hurricane season. The

result was a dry spell that hit particularly hard in the western Amazon, a region that normally has more rainfall than other parts of the forest.

"It caught people by surprise," says William Laurance, an ecologist with the

Smithsonian Tropical Research Institute, based in Balboa, Panama. "We hadn't seen a drought like that before."

The media quickly blamed global warming, but climate researchers warn that it may not be possible to make such



How will the rainforest respond to drought?

A. LIMA/EMPHICS/AP



Oskar Kaibyshev (centre) and his lawyers: other scientists have received much harsher sentences.

high treason, and there may be more unpublished cases, says Chudnovsky.

Since the collapse of the Soviet Union, science in Russia has suffered financially. Foreign grants and collaborations have helped keep many institutes alive. But accusations against scientists who work with foreign groups or companies have also become widespread.

The threat of prosecution is making things tense for Russian scientists who maintain international contacts and aim to commercialize their research, says Sarah Olmstead of

the human-rights programme at the American Association for the Advancement of Science.

Prosecutors in Kaibyshev's case had called for an unsuspended six-year sentence. The relatively mild verdict could be due, some say, to international outrage over the case.

"It's a big relief to us all, but one case doesn't make a statistic," says Chudnovsky. "I'm still waiting for an imprisoned scientist to be released before I can believe that things are getting better."

Quirin Schiermeier

a link — in part because vast areas of Amazonia have no weather stations.

But early results from studies suggest that the Amazon could have more such events in the future. In a paper to be finalized this month, researchers from the United Kingdom and Brazil use a climate model developed at the Hadley Centre for Climate Prediction and Research in Exeter, UK, to study whether the 2005 event was a one-off or a taste of things to come.

The team says the model predicts the latter. "The 2005 situation will be more frequent by 2050," says co-author Jose Marengo of the Center for Weather and Climate Prediction in Cachoeira Paulista, Brazil.

Last year's drought also bolsters a controversial 2004 finding about the future of the Amazon rainforest. Peter Cox of the Centre for Ecology and

Hydrology in Winfrith, UK, who is also an author of the new paper, had predicted that more frequent droughts would wipe out around 65% of the Amazon forest cover by 2090 (P. M. Cox *et al. Theor. Appl. Climatol.* 78, 137-156; 2004). Many questioned the result when it was published, in part because some of the droughts predicted by the model were caused by warm waters in the Atlantic — a phenomenon that hadn't been observed at the time. Now that such an event has been seen, says Laurance, researchers are looking at Cox's results with a little less scepticism.

Other studies disagree, predicting less drought and forest destruction than Cox's model. But more and more evidence suggests that drought will take a serious toll on the Amazonian forest.

Daniel Nepstad, an ecologist at the Woods Hole Research

Center in Massachusetts, has been simulating drought since 1998 with a plastic canopy covering 10,000 square metres of forest in the eastern Amazon. His results show that during drought, more large trees die off, all trees produce less wood, and there are more fires. In a paper under review at *Ecology*, Nepstad calculates that the combination of dead trees and reduced wood formation in Amazonia as a whole led to an extra half-a-billion tonnes of atmospheric carbon in 2005. In contrast, the Kyoto Protocol on Climate Change aims to reduce global carbon emissions by a billion tonnes annually by 2012.

Nepstad is now lighting test fires to assess whether one fire makes the forest more vulnerable to others. "Our biggest fear is that the forest will be invaded by highly flammable grasses," he says. ■ Jim Giles



autoMACS™ Pro Separator

Cell sorting of up to six samples in a standardized procedure

- MACS® MiniSampler for sequential cell sorting
- Different cooling tube racks for excellent viability
- Optimized, sensor-controlled process for less hands-on time
- High-speed sorting of up to 10^7 cells per second for fast results
- autoMACS™ Platform renowned for reliable results from any species

HIGH-QUALITY MAGNETIC CELL SEPARATION – NOW EVEN EASIER!



Unless otherwise specifically indicated, all Miltenyi Biotec products and services are for research use only and not for therapeutic or diagnostic use. MACS is a registered trademark of Miltenyi Biotec GmbH. autoMACS is a trademark of Miltenyi Biotec GmbH.

Miltenyi Biotec

Miltenyi Biotec GmbH Phone +49 2204 8336-0
Friedrich-Ebert-Str. 68 Fax +49 2204 85197
51429 Bergisch Gladbach macs@miltenyibiotec.de
Germany www.miltenyibiotec.com


AIRPORT SECURITY IN THE SPOTLIGHT

News@nature.com gathers our stories on security technologies.

www.nature.com/news

G. RIZER/AP

Florida lures research institutes east

Florida is on a drive to recruit Californian research institutes, dangling tempting financial offers to persuade renowned laboratories to set up branches in the state.

At least three research institutes are jockeying for a share of the \$245 million recently allocated to job-creating projects in Florida. The organizations whose leaders are being wined, dined and flown to the state on private jets include the Burnham Institute, a biomedical research facility based in La Jolla, California; SRI International, a non-profit research institute in Menlo Park, known for developing Silicon Valley technologies; and the Torrey Pines Institute for Molecular Studies, also based in La Jolla.

The driving force behind the recruitment is Florida Governor Jeb Bush, a Republican and the brother of President George W. Bush. Earlier this year, Governor Bush got \$245 million from the Florida legislature to create high-tech industry jobs.

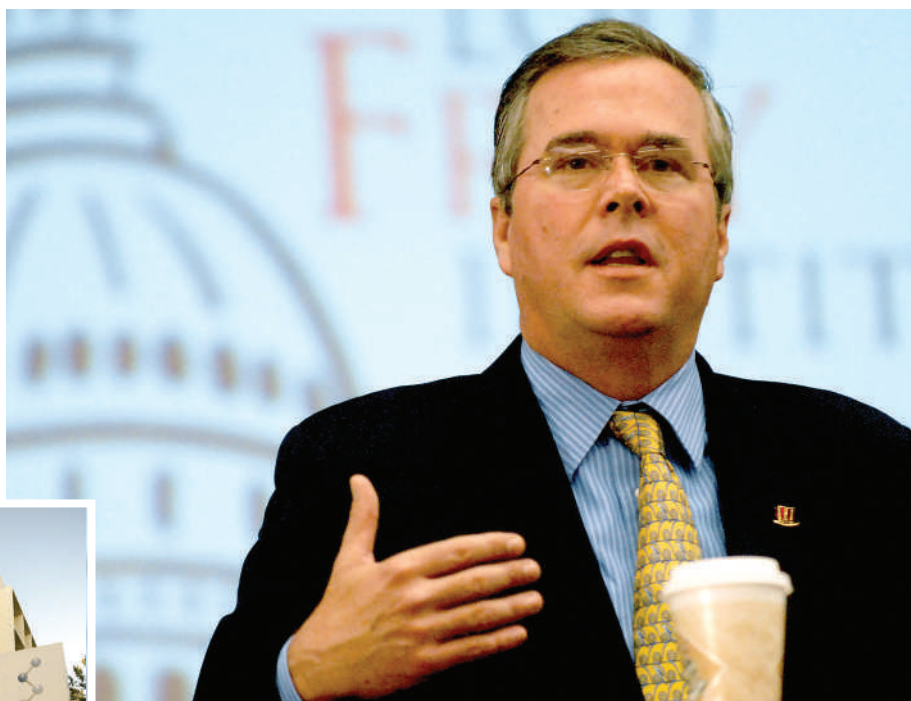
He is now using that money to woo research leaders from California, the state apparently most interested in what Florida has to offer. A state legislative committee is scheduled to meet on 17 August to consider proposals on how to divvy up the money.

Under Governor Bush's proposals, local governments will partner with a research institute and offer land and buildings worth nearly \$100 million. Each partnership will then ask the state to match that investment.

A 2003 agreement between Governor Bush, the county of Palm Beach, and the California-based Scripps Research Institute, to build a research complex with the Scripps name, provides a model for how this could work. So far, more than \$90 million has been spent on 'Scripps Florida', although no permanent buildings have yet been constructed.

Separately, Scripps has launched an initiative to build a leading centre for avian influenza research in Florida. Called Project Checkmate, the plan calls for Scripps to join forces with IBM's Thomas J. Watson Research Center in Yorktown Heights, New York. Researchers from both institutes would use supercomputers and a biosecure laboratory facility to try to predict each year's avian-flu mutations in advance.

"Biotechnology combined with large-scale computing could really move both fields for-



AP PHOTO/P. M. EBERHACK

ZUMA PRESS/NEWS.COM



Grand plan: governor Jeb Bush plans to create jobs in Florida by persuading California-based research institutes, such as the Scripps Institute (left), to set up branches there.

ward," says Ajay Royyuru, head of computational biology at IBM's Watson Center. Project leaders are seeking \$520 million over five years from the US National Institutes of Health, as well as \$18 million from Florida.

Many in Florida support such collaborations, saying they will bring much-needed jobs to the state. But not everyone is happy. "These are absolutely horrible deals, giving away the store," says Dan Gelber, a Democratic state representative from Miami Beach. Joseph Cortright, an Oregon-based economist who studies high-tech industries, agrees that pouring money into new facilities doesn't necessarily create better jobs or broad economic benefit to the region.

Gold rush

Some Florida universities also aren't thrilled. "The state needs to support research institutions that are already operating," says Abdul Rao, a vice-president for research at the University of South Florida. To help pay for the transfers, Governor Bush is proposing shifting millions of dollars from state-university budgets.

The president of the Burnham Institute, molecular biologist John Reed, says money was

the principal reason the Burnham wants to add a Florida facility. In recent years the institute had asked California legislators for additional funding, but without success. Reed wouldn't comment on financial specifics of the deal, but the Burnham reportedly wants about \$150 million of the state's money.

Expanding into Florida, Reed argues, would raise the Burnham's profile. "I think it is a great advantage, having a presence at the national level," he says. The institute has narrowed possible locations for its expansion to the Orlando area and to Port St. Lucie on the Atlantic coast.

Meanwhile, SRI International has its eye on creating a water-research facility near Tampa, in conjunction with the University of South Florida. Project costs weren't available, and SRI officials declined to comment.

As for the Torrey Pines Institute, it plans to partner with governments in either Boca Raton or elsewhere in Palm Beach County and hopes to eventually have 20 principal investigators there. The institute was personally recruited by Governor Bush, says chemist Richard Houghten, its president and founder. "It's very exciting," he says.

Rex Dalton

With additional reporting by Emma Marris

SPECIAL REPORT

The methane mystery

The claim that living plants emit the greenhouse gas methane has shaken up atmospheric scientists. **Quirin Schiermeier** talks to the experts trying to make sense of the measurements.

It took 18 years, but Paul Crutzen and Eugenio Sanhueza finally found a use for their data. In 1988, the two atmospheric chemists had discovered large amounts of the greenhouse gas methane building up at night over Venezuela. This March, they dusted off their data and used them to calculate a surprising number: that the world's tropical savannahs may produce 30 million to 60 million tonnes of methane each year¹.

Remarkably, that number is similar to one reported in January, when a German chemist announced that living plants give off methane². His claim rattled many, because textbooks hold that methane is produced from organic matter decaying in oxygen-free environments, not from living plants. If true, his finding could account for a substantial fraction of the methane entering the atmosphere — potentially throwing off calculations of how much humans contribute.

In the past seven months, atmospheric scientists have scrutinized the discovery, and they're finding that methane does not yield up its secrets easily. Some, like Crutzen, find the work convincing. "We could have seen the effect a long time ago, but we missed the boat,"

says Crutzen, who works at the Max Planck Institute (MPI) for Chemistry in Mainz, Germany, and shared the 1995 Nobel Prize in Chemistry for his work on atmospheric ozone. Sanhueza works at the Venezuelan Institute for Scientific Investigations in Caracas.

Others are not swayed. "I am not yet convinced that the effect is real," says David Beerling, a palaeoclimatologist at the University of Sheffield, UK. "The experiments leave a lot of questions open, and the findings still await independent verification."

And some just aren't sure. "The new source is very hard to distinguish from known methane sources such as wetlands, and the scarce existing data can neither prove nor disprove its existence," says Martin Heimann, an atmospheric chemist at the MPI for Biogeochemistry in Jena, Germany.

Time to adjust

The debate began with the work of Frank Keppler, a geochemist at the MPI for Nuclear Physics in Heidelberg, Germany, and his colleagues. The researchers grew various plants in isolated chambers, then measured the background methane concentrations. The fluxes were tiny; but when extrapolated to global vegetation they pointed to the existence of a huge, overlooked source of methane. Plants, it seemed, could account for up to 40% of total methane emissions (see 'Missing mechanism').

As a greenhouse gas, methane traps heat 20 times more efficiently than carbon dioxide, yet its sources and sinks are less well understood. Total methane emissions are estimated to be around 550 million tonnes per year. The known sources include wetlands, rice farming, the stomachs of ruminant animals, biomass burning, landfills and energy generation. And now, perhaps, plants.

If Keppler is right, rapid tropical deforestation could explain why atmospheric methane concentrations stopped rising in the 1980s, after decades of increase. Looking further back in time, the effect could have played a role in the transition from glacial periods to warmer climates, says Thomas Röckmann, an atmosphere researcher at the Utrecht University in the Netherlands. Warmer temperatures could mean more plants, which in turn produce more heat-trapping methane.

But most researchers are cautious about



jumping to conclusions, particularly in extrapolating Keppler's laboratory observations to global values. Keppler calculated a global source in the range of 62 million–236 million tonnes per year. Other ways of scaling up his data suggest the source might be substantially smaller.

For instance, critics say Keppler overestimated the amount of methane-producing plant tissue by basing his calculations on global 'net primary production' of vegetation, which includes roots and stems, rather than global leaf mass only. When climate modeller Miko Kirschbaum of the Cooperative Research Center for Greenhouse Accounting at the Australian National University in Canberra adjusted the calculation accordingly, he arrived at global emissions of only 10 million–60 million tonnes per year³.

Others have looked at the carbon-isotope ratios of methane in ice cores. Different sources of methane yield different isotope signatures. A 2,000-year ice-core record from

Missing mechanism

How plants might produce methane remains a mystery.

Frank Keppler, the chemist who proposed the idea in January, ruled out the possibility that the gas could have come from bacteria in leaves, or enzymes in plant cells. This suggests that plants somehow produce the methane 'autonomously', without a protein to catalyse the reaction.

Instead, Keppler proposed a chemical pathway involving pectin. These molecules, a mix of different sugars used by plant cells, are constantly being synthesized and degraded — a process that could produce methane. But how that would work is the subject of much debate.

Other chemical reactions that could release methane would require the presence of unstable carbohydrate molecules in the plant cells, says botanist Jürgen Kesselmeier of the Max Planck Institute of Chemistry in Mainz, Germany. Plant physiologists don't yet know what those might be.

Q.S.

**ANTARCTIC CLIMATE**

Missing snow in weather 'hindcast' could mean sea-level rises will be worse than was feared.

www.nature.com/news



Green greenhouse? If plants produce a lot of methane, it could affect political efforts to cut emissions.

Antarctica yielded an upper limit of 46 million tonnes per year⁴.

But not everyone is revising Keppler's findings downward. One new simulation, of how methane moves through the atmosphere, suggests that plants yield on the order of 125 million tonnes per year. Sander Houweling, of Utrecht University, simulated methane fluxes both with and without methane emissions from plants. He found that the results better match observations if vegetation is included as a source of methane⁵.

Keppler, a co-author of this study, says he can "live very well" with attempts to re-interpret and downscale his findings. The 39-year-old researcher recently received a European Young Investigator Award, allowing him to continue work in the area for at least five more years.

"Unpublished data show that some plant species emit up to 4,000 times more methane than others."

And he promises further surprises. Unpublished data from experiments carried out last year in Brazil, on five randomly selected plant species, show that some species emit up to 4,000 times more methane than others. If this is true, it will make it even more difficult to calculate the size of the global source, he says.

If plants are giving off more methane than was thought, then another source of methane must have been overestimated. Most scientists suspect that this is wetlands. More unlikely, but politically more contentious, is the possibility that one or more of the human-related sources might be much smaller than was thought, says Johannes Lelieveld, a director of the MPI for Chemistry.

The German gas company Ruhrgas has already asked Lelieveld whether it can revise

downward its estimates of how much methane its power plants have emitted. "That's probably going too far," says Lelieveld. "But a major discovery like this does require reassessment of greenhouse-gas sources."

As usual, more data would help. "The ultimate test is to do more and better measurements," says Ed Dlugokencky, an atmospheric chemist with the US National Oceanic and Atmospheric Administration in Boulder, Colorado, who oversees the methane part of a global air-sampling network.

Dlugokencky is convinced that the effect is real. "I'm not surprised we missed it," he says. "Our network of 60 stations is just too small to be able to detect it, and we are particularly weak in the tropics where it seems to be most pronounced."

In flux

Another way to hunt methane could be from space. As reported last year, the European Envisat satellite has detected large plumes of methane over tropical rainforests⁶. But remote sensing cannot prove that the gas stems from trees, as opposed to swamps, bogs or wetlands underneath.

"You can learn a lot from space," say Crutzen. "But to verify sources you need to go back to places such as Venezuela and Brazil and look at methane *in situ*."

Eddy-flux towers, which are already being used at more than 200 sites worldwide to measure the movement of carbon dioxide in and out of the atmosphere, might help. Methane is less abundant than carbon dioxide, and so harder to monitor. But new and more powerful lasers will measure gas concentrations more accurately. Soon, says Heimann, it may be possible to measure methane fluxes above tree canopies in real time — perhaps even tracing them back to individual plants.

Such information should improve estimates of the global methane budget. With that, researchers hope to move closer to their ultimate goal: developing a computer model that accurately reproduces past climate and all its biospheric feedbacks.

A model of the past should also provide projections for the future. "You need to understand the entire greenhouse budget," says Dlugokencky, "before you can start thinking about mitigating climate change." ■

1. Crutzen, P. J., Sanhueza, E. & Brenninkmeijer, C. A. M. *Atmos. Chem. Phys. Disc.* **6**, 3093–3097 (2006).
2. Keppler, F. *et al. Nature* **439**, 187–191 (2006).
3. Kirschbaum, M. U. F. *et al. Funct. Plant Biol.* **33**, 521–530 (2006).
4. Ferretti, D. F. *et al. Atmos. Chem. Phys. Disc.* **6**, 5867–5875 (2006).
5. Houweling, S. *et al. Geophys. Res. Lett.* (in the press).
6. Frankenberg, C. *et al. Science* **308**, 1010–1014 (2005).

P. ADAMS/GETTY IMAGES

NASA climbs back on board dark-energy mission

NASA is renewing its support for a space mission to investigate dark energy, the mysterious force that could make up two-thirds of the energy density of the Universe.

The Supernova Acceleration Probe (SNAP) was first proposed by scientists at the US Department of Energy's Lawrence Berkeley National Laboratory in California in 1999. In 2003, NASA joined forces on the project, but pulled out a year later when President George W. Bush announced plans to redirect NASA towards exploration of the Moon and Mars (*Nature* 427, 667; 2004).

On 9 August NASA said it will re-enter SNAP by supporting a concept study, which will also provide a cost estimate, likely to be several hundred million dollars. "I'm optimistic," says Saul Perlmutter at Lawrence Berkeley Laboratory, a principal investigator on the SNAP collaboration.

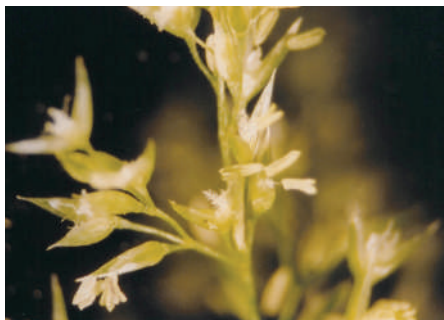
But the die is not yet cast. NASA also announced plans to study concepts for two smaller missions to study dark energy: ADEPT from Johns Hopkins University in Baltimore, Maryland, and DESTINY from the National Optical Astronomy Observatory in Tucson, Arizona.

Transgenic grass strain escapes into the wild

A transgenic strain of grass bred for golf courses has spread from test plots and established itself in the wild, researchers say. Ecologists warn that the hardy, fast-spreading strain could overwhelm native species in the northwestern United States.

Scientists working for the US Environmental Protection Agency in Corvallis, Oregon, say they have found the plant, known as creeping bentgrass (*Agrostis stolonifera*), 3.8 kilometres away from experimental plots. Seeds and pollen have been dispersed by the wind, they report in a forthcoming issue of *Molecular Ecology*.

The strain, bred by horticulture supplier Scotts, in Marysville, Ohio, is engineered to be resistant to the herbicide Roundup. This



The grass is always greener: a transgenic strain of creeping bentgrass has escaped from test plots.

Student's organic design gives lab a cell structure

It's an opportunity any architect would die for. But it fell to Sloan Kulper, an architecture student at the Massachusetts Institute of Technology (MIT), whose interest in biology led him to sit in on the lectures of Shuguang Zhang, an MIT biomedical engineer.

When Zhang was asked to advise on a new nanobiomedical institute at his alma mater, Sichuan University, he encouraged Kulper to design a building for it, drawing on the biology he had learned. The resulting cell-shaped building has passed technical review. Its interior is packed with biological references, from a crystal-shaped lecture theatre to handrails jointed with nucleosome shapes. Zhang says the building would be too expensive in Europe or the United States, but will cost no more than US\$10 million in China. He is still seeking funds.



S. ZHANG

will make it hard to eradicate from areas where other grasses are grown and managed with herbicides, warn ecologists.

Animal-rights groups force researcher to quit studies

Chalk one up for the animal-rights activists. In a 4 August e-mail to activists titled "You win", Dario Ringach, a neurophysiologist at the University of California, Los Angeles (UCLA), announced that he would cease all research on animals immediately. "Please don't bother my family any more," he added.

Ringach had been harried by activists protesting against his work on visual processing in macaques. He gave up after an incident on 30 June in which Lynn Fairbanks, another primate researcher at UCLA, found an unexploded home-made bomb on her front porch.

UCLA has defended the animal research it hosts, arguing that it helps to improve human health and society.

But Jean Barnes, head of the Primate Freedom Project of Fayetteville, Georgia, is not satisfied. Barnes refuses to take Ringach's home address off the Primate Freedom Project website until he apologizes.

Gates gift to Global Fund shames governments

On the eve of the XVI International AIDS Conference, held in Toronto this week, the Bill and Melinda Gates Foundation stepped up pressure on governments by pledging the largest non-governmental donation yet to the Geneva-based Global Fund to Fight AIDS, Tuberculosis and Malaria.

The Gates Foundation said on 9 August that it will give US\$500 million over five years on top of the \$150 million it has already donated to the fund. Since 2002,

the Global Fund has provided \$5.5 billion, mostly from governments, to 132 countries.

The fund's executive director, Richard Feachem, says it still needs \$500 million to finance its next round of grants in November. Announcements of new commitments to the Global Fund may come during the AIDS meeting. Paul Zeitz, executive director of the non-profit organization Global AIDS Alliance, based in Washington DC, estimates that the Global Fund will require \$11 billion per year by 2010.

Electronic ark to hold all known animal species

It's being called a 'genome project' for zoology, the field's equivalent to the Moon landing. ZooBank, an online database of all known animal species, was launched on 10 August (www.zoobank.org).

The database was established to end the confusion caused by information on new species being published in different journals (see *Nature* 437, 477; 2005). It already contains details of the 1.5 million species compiled as the electronic archive of the past 150 years of the *Zoological Record*. With the backing of the International Commission on Zoological Nomenclature (ICZN), ZooBank will be the official registry for new species.

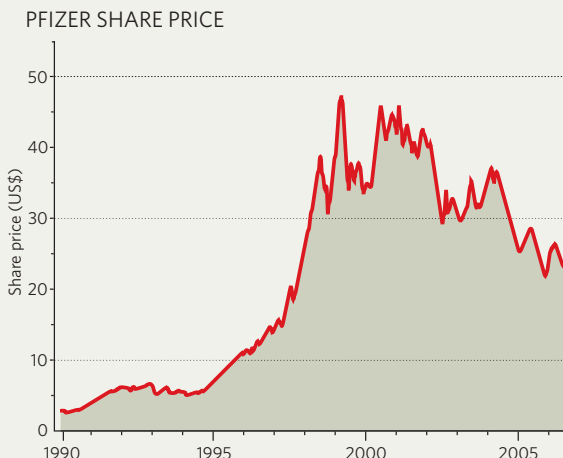
"ZooBank will revolutionize the way in which those people using scientific animal names can work," says Andrew Polaszek, a taxonomist at the Natural History Museum in London and executive secretary of the ICZN.

Correction

The News story "Wildlife caught in crossfire of US immigration battle" (*Nature* 442, 338–339; 2006) incorrectly ascribed the research on counting illegal immigrant trails in the Buenos Aires National Wildlife Refuge to Peter Morrison and the Pacific Biodiversity Institute. The research was actually done by the refuge's Dan Cohan.

BUSINESS

C. EAST/REUTERS



H. RAY ABRAMS/AP

More fizz at Pfizer?

Young blood is taking the helm at the world's biggest drug firm. **Colin Macilwain** assesses whether this is likely to revive its share price.

As chief executive, Hank McKinnell cut an avuncular figure in his company's television commercials, reassuring viewers that when their doctors needed medicine, Pfizer would be there.

But when the board of the world's largest drug firm brought McKinnell's five-year tenure to an abrupt close last month, few on Wall Street were complaining. The fact that his replacement — Jeff Kindler, a 51-year-old lawyer — is a relative newcomer to the firm was loudly applauded by observers who want to see Pfizer ring the changes.

"They are trying to send a message that they are willing to look at something different," says Al Rauch, an analyst at stockbroker A G Edwards in St Louis, Missouri. Pfizer, he says, has "a very consensus-driven corporate culture that is not always conducive to innovation".

But despite the generally upbeat reaction to Kindler's appointment, the value of Pfizer's stock has remained unmoved. And close observers of the company question whether the appointment will do much to change its outlook.

Like other major pharmaceutical firms, Pfizer is in a hole. Patents will soon run out on its major products, including the world's best-selling drug, cholesterol treatment Lipitor (atorvastatin). "They need to generate \$10 billion in extra revenue each year and that's impossible," says Peter Rost, a former vice-president of Pfizer who runs a carping blog, <http://peterrost.blogspot.com>, on its predicament. "The only way they can stay afloat is as a black hole, sucking in other companies to boost revenue. I think they are going

to collapse under their own weight."

Pfizer's \$190-billion market capitalization suggests that this prognosis is a little harsh. And close observers of the drug industry say that the New York-based company has done better than most of its peers in responding to the strategic challenges that the sector faces.

"In some ways, Pfizer has led the industry," says Martyn Link, a pharmaceuticals analyst at Wood Mackenzie in Edinburgh, UK. He notes as examples the cost-cutting programme launched by McKinnell last year to achieve annual savings of \$4 billion by 2008; the successful offload in June of its consumer drugs business to Johnson & Johnson of New Jersey; and an aggressive strategy of purchasing hot biotechnology companies.

Strategic switch

The drive for biotech acquisitions, which aims to buy two or three substantial operations each year for between \$1 billion and \$3 billion, has seen Pfizer get its hands on stars such as California-based antibody specialist Bioren and Vicuron Pharmaceuticals, a Pennsylvania company that is developing anti-infective drugs.

But the company hasn't convinced investors of its successes, Link says. He argues that Pfizer has a number of potentially lucrative products that have recently entered the market — notably the cancer treatment Sutent (sunitinib) and Exubera, an inhalable form of insulin. "They need to be able to sell their story to investors," he says.

McKinnell, a Canadian business-school graduate who has spent 35 years at Pfizer, was

Bitter pill: disappointing results have seen Hank McKinnell (left) removed as Pfizer's chief executive to make way for Jeff Kindler.

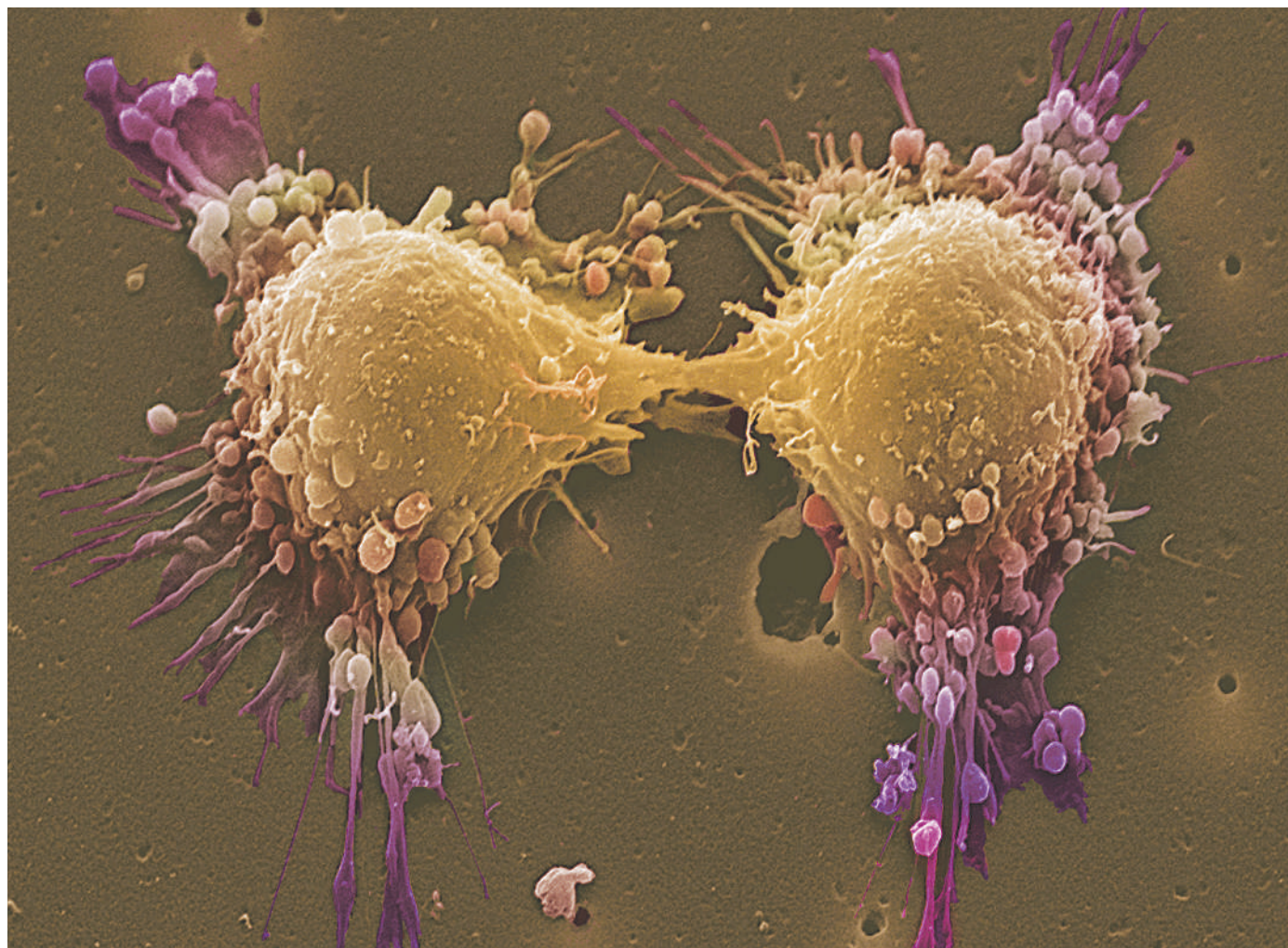
responsible for much of this progress. But with the share price in steady decline since his appointment, the board announced on 28 July that he would retire a year earlier than planned, and that Kindler, the company's chief lawyer for the past four years, would take his place.

The manner of the appointment has left blood on the walls. Kindler's two rivals for the position, Karen Katen, head of the firm's dominant pharmaceutical division, and David Shedlarz, its finance chief, haven't been offered any of the usual sops, such as board positions. US press reports suggest that Kindler's early efforts to reach out to the duo have fallen on deaf ears, and there is speculation that they will leave. The appointment of a lawyer whose main operational business experience was at McDonald's is also unlikely to inspire Pfizer's increasingly restive research staff.

Link is sanguine about that. "Pfizer's culture is accepting to people of quality and ambition," he says. But he adds that there isn't much Kindler can do in the short term to speed McKinnell's reforms. "There really aren't a lot of strategic options that Pfizer hasn't already tried," he says.

One area of strength for Kindler is patent law: he has been closely involved in critical court battles to preserve exclusive rights. But elsewhere, his room for manoeuvre may be restricted. "It's going to be difficult to change the corporate culture," says Rauch, adding that attempts to do so "are likely to meet resistance".

For investors, the reality is that they will have to get used to slower growth, and perhaps more modest margins, from major drug firms. "Investors came to expect double-digit growth every year," Link says. "Now that's unrealistic. It is a much more challenging environment." ■



S. GSCHMEISSNER/SPL

CANCER

New fronts in an old war

Like so many wars, the war on cancer can cause despair, with final victory sometimes seeming inconceivable. In the 35 years since President Richard Nixon declared his battle on everybody's enemy, the immense progress in scientific understanding of the disease has yet to deliver the crushing blow that was hoped for.

Indeed, the problem seems to have become less tractable, not more so. The concept of cancer as a single disease has been replaced by a view of it as many different ones, each with its own molecular process, each requiring its own therapeutic approach. Thanks to advances in molecular biology, therapies targeted at specific disease processes are continually being developed. But many believe they are likely to roll back the cancer mortality figures slowly, rather than making a quick and dramatic impact.

If we leave the military analogies aside, we can see that there are possibilities for progress — not for final victories, but for saving lives in the fairly near future, and developing new approaches to stubborn problems. This special section of *Nature* looks at three of these new angles on cancer research and what they promise to contribute.

One area where science could improve survival rates is by finding ways to detect cancers early. As Laura Spinney explains on page 736, existing

treatments are far more effective when delivered early, and technologies to detect disease earlier than ever before are becoming available. The problem is choosing the right signs and understanding what they mean.

Even if we catch cancers early, are the drugs targeting them correctly? A new theory suggests that they could be missing something, says Alison Abbott (see page 742). A small subset of cells, called cancer stem cells, could re-seed tumours after treatment. Researchers are trying to develop drugs aimed at cancer stem cells, in the hope of preventing the recurrence of tumours — the leading cause of cancer deaths.

New drugs need to be tested. One long-standing criticism of cancer research is that about nine out of ten drugs that look good in preclinical tests fail to fulfil that promise in human trials. Many point the finger at the mouse, the stalwart animal model of cancer, for being too mouse-like to be a good model for humans. But that is changing, says Carina Dennis (see page 739), as researchers find ever-more ingenious ways to engineer mice to be more like us. ■

For more on cancer see the Editorial on page 720, and visit our web focus at www.nature.com/nature/focus/cancerhorizons

Caught in time

The detection of cancer at an early stage in its development can be life-saving. With research efforts under way to find better methods to detect minuscule tumours, **Laura Spinney** finds out how near some of these cancer 'biomarkers' are to the clinic.

When it comes to halting cancer, nobody disputes the value of early detection. Dr George Papanicolaou's smear test, which measures changes in the cervical lining, is a testament to the importance of screening. Since 1950, when it was introduced, the Pap smear has reduced mortality due to cervical cancer by 70% in developed countries. About half of lung-cancer patients whose cancer is caught while it is still restricted to the lung survive. Nearly all of those diagnosed after it has spread die. (See graphic, opposite.)

Yet for all its promise, early detection remains the poor relation of cancer treatment. In 2005, the US National Cancer Institute (NCI) spent around 8% of its research budget on detection and diagnosis, compared with 22% on treatment. Britain's National Cancer Research Institute, a collaboration of the country's major cancer-research funders, spent 9% of its 2004 budget on early detection, diagnosis and prognosis, versus 19% on treatment.

Seeing the signs

Nevertheless, the technology needed to detect cancers at the very earliest stages is within sight. Advances in mass spectroscopy and in the detection of mutant DNA, for example, are now making it possible to pick up very low levels of tell-tale molecules. Meanwhile, medical-imaging technology is improving our ability to peer inside the body (see 'Powers of detection', overleaf). Ultimately, scientists hope to devise straightforward, non-invasive tests, such as screening blood samples, to pick up on the earliest signs of problems — perhaps even before cells become fully cancerous.

As a first step, scientists are hunting for effective 'biomarkers' — DNA or proteins that act as indicators of normal or pathological biological processes that could be used to screen for the presence of a cancer. But a big challenge facing biomarker hunters is that blood is a complex soup of molecules containing everything from

the ubiquitous protein albumin to smaller molecules present only at trillionths of albumin's levels. As well as posing a challenge for molecule-detection technologies, this makes it even harder for biologists to pick out which proteins are relevant.

"The important markers are likely to be in the low-abundance range, so the problem is searching for the needle in the haystack," says Lee Hartwell, Nobel laureate and director of the Fred Hutchinson Cancer Research Center in Seattle, Washington. Yet in the past few years, the technology for identifying potential biomarkers has advanced quickly, and many have been

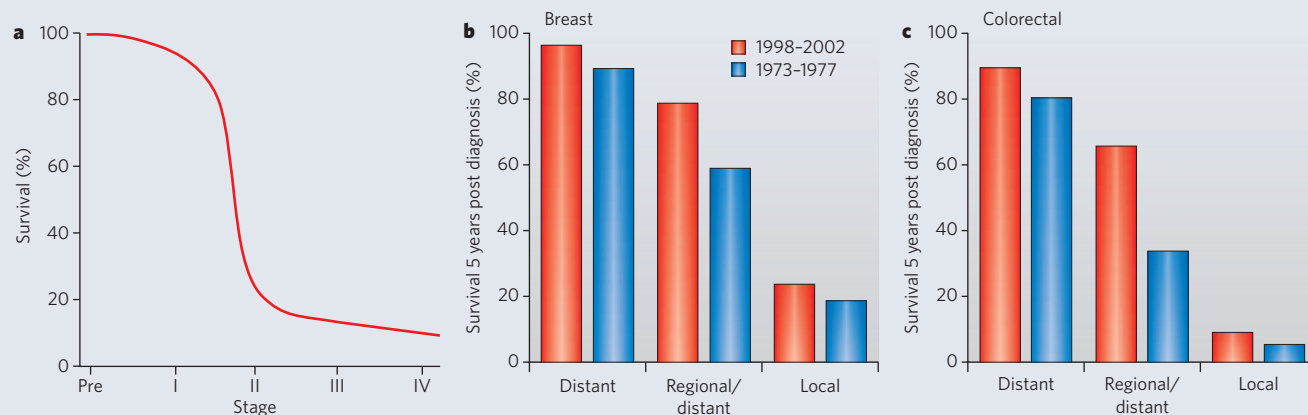
added to the list. For example, modern mass spectrometers provide high enough resolution that mass spectra of small proteins can now be obtained and matched against sequence databases for identification.

Another method, called high-throughput proteomic profiling, relies on large arrays of different antibodies that specifically bind to proteins found in the blood. These can detect new proteins and changes in the levels of existing proteins that correlate with the development of cancer. Sam Hanash's group at Fred Hutchinson has been using proteomic profiling to identify markers for early-stage lung cancer, using two strategies¹. The first is to look for proteins that have been shed by the



S. PEMBERTON

THE VALUE OF EARLY SCREENING



a, Catch a tumour before it becomes malignant (pre-malignant stage) or while it is still small and locally confined (stage I) and the chances of a cure are high. They drop sharply once the cancer spreads to local tissues and lymph nodes (II-III) or other organs (IV). **b**, **c**, Despite improvements in treatments since the 1970s, patients still have a greater chance of surviving if cancers are detected when they are locally confined (local) and not spread to neighbouring tissue (regional) or other organs (distant).

tumour and are circulating in the bloodstream; the second is to search for an immune response to the tumour. The host immune system may not be able to destroy the tumour cells, but it probably recognizes them as problem cells, and Hanash wants to harness that raised antibody response as a diagnostic tool. "We're finding that some circulating proteins are inducing an immune response," he says. "We're also finding antibodies to circulating proteins."

Hanash's biomarkers are just some of many that scientists have identified. Underpinning all this work is the understanding that using a single biomarker to diagnose a cancer is unlikely to suffice. "We're all agreed now that to rely on one marker to make a diagnosis is not going to work," says Hanash. No single biomarker can detect a given cancer with 100% sensitivity (meaning that all diseased subjects would test positive) and 100% specificity (with all healthy subjects testing negative). Panels of biomarkers with different individual sensitivities and specificities are therefore needed.

A team led by David Sidransky at Johns Hopkins University School of Medicine in Baltimore, Maryland, for example, is trying to find ways of improving the accuracy of the blood tests used to detect a protein called prostate-specific antigen, or PSA. PSA is produced by prostate cancers, but can also result from benign prostate enlargement or inflammation. Sidransky's team has identified an alteration to a gene called *GSTP1* that is unique to prostate cancer cells. The hope is that men with high levels of PSA in their blood could be referred for a biopsy to test for this genetic alteration, to reduce the number of false positives that result from relying on PSA levels alone².

To become fully cancerous, cells go through a complex, stepwise progression from normal cells to diseased ones. Scientists argue³ that, for optimal early detection, researchers need to find biomarkers targeted to a stage in this progression where changes can be detected with the highest sensitivity and specificity. That may

be a 'pre-malignant' phase, when cells begin to behave abnormally, but have not yet developed all the characteristics of tumour cells.

One example of where this approach is being applied is in research on the common and sometimes crippling gynaecological condition endometriosis, where tissue that normally lines the womb appears on other pelvic organs. Women with endometriosis have a higher risk of developing a particular type of ovarian cancer and last year, Daniela Dinulescu, now at Brigham and Women's Hospital, Harvard Medical School, discovered genetic pathways underlying the progression from endometriosis to cancer in mice⁴. This raises the possibility that a test could one day pinpoint early genetic changes in women. In separate work, Dinulescu is collaborating with Hanash to validate around 20 biomarkers for early ovarian cancer.

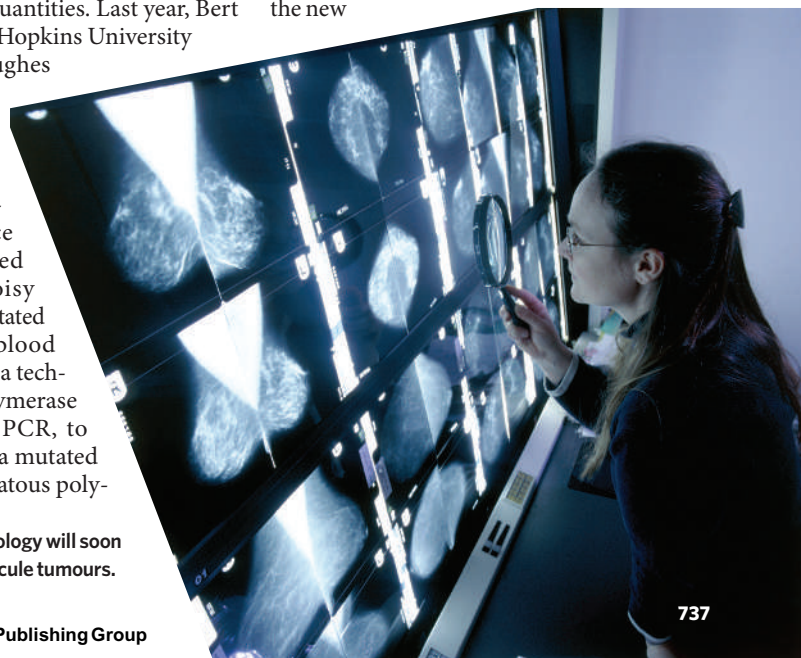
Risk assessment

As researchers rush to assemble biomarker panels, the limiting factor to early detection is becoming the ability to detect these biomarkers in ever smaller quantities. Last year, Bert Vogelstein of Johns Hopkins University and the Howard Hughes Medical Institute (HHMI) in Chevy Chase, Maryland, showed for the first time that it is possible to detect trace amounts of mutated DNA against a noisy background of unmutated DNA, in a simple blood sample⁵. He adapted a technique called the polymerase chain reaction, or PCR, to detect fragments of a mutated gene called adenomatous poly-

posis coli (APC) in the blood of patients with advanced colorectal cancer. PCR allows researchers to selectively amplify a particular stretch of DNA using an enzyme, which can start work on very small quantities of DNA. Vogelstein's team also detected mutant APC molecules in more than 60% of patients with early, curable colorectal cancer; these were circulating in the patients' blood in extremely low quantities.

Bioengineer Steve Quake of the HHMI and Stanford University thinks that this is just the beginning. "The fundamental limit of detecting single molecules has been reached in many cases of scientific interest, so in principle, this should be applicable to cancer detection as well," he says.

Given that there is no such thing as a single magic marker for any cancer, Quake feels that the ideal way to detect early disease would be to sequence and so identify all the hundreds of thousands of molecules in a blood sample. "As crazy as it sounds, this is something that is coming within reach of the new



Medical imaging technology will soon be able to detect minuscule tumours.

Powers of detection

Many scientists working on the early detection of cancer hope to find molecules, or biomarkers, in a person's blood that betray the presence of a tumour at the earliest stages of its development. But a blood test will simply tell you that a tumour exists, not where it is.

To pinpoint a tumour's location, researchers are developing medical-imaging technology that is much more sensitive than conventional mammography or computed tomography (CT) scans and is able to pick up tiny tumours.

The key to determining a tumour's location is to inject a compound or probe, visible to imaging machines, into the bloodstream that seeks out a particular biomarker in

the tumour. The compound will then accumulate in the tumour, making it light up on scans. With conventional positron emission tomography (PET) scans, where the probe is radioactive, there are limits to how much of it you can safely introduce into a person. But with newer techniques, such as optical (pictured below) or ultrasound imaging, scientists can inject more of the probe to boost the signal, as the probes, for example, antibodies attached to fluorescent chemicals, are less toxic.

Plus, says Sam Gambhir, director of Stanford University's Molecular Imaging Program, the properties used to assess the signal, such as wavelength, scatter and absorption,

simply allow you to see more.

Gambhir says it is currently possible to see down to the level of thousands of cells in animal models such as mice. In humans, the increased complexity and depth of our tissues means the limit is closer to millions of cells. Seeing a single cell in humans would depend on developing strategies with much higher signal amplification and very low noise. "Nothing we have right now would allow that," he says, "But it is certainly feasible that in humans we could get down to thousands of cells. The imaging community is working on that."

A 1 mm³ tumour contains about 3 million cells; if scientists could

get down to thousands of cells, they could detect tumours around 0.1 mm³ in size. Most researchers think that detecting tumours in the range of 0.5 mm³ to 1 mm³ is achievable within the next five to ten years, says Gambhir.

In part, he adds, such improvements depend on selecting the right biomarkers. With the US-based Canary Foundation, a non-profit organization dedicated to the early detection of cancer, Gambhir has brought together members of the imaging community with those working on biomarkers to ensure the right ones are developed. It's an exciting collaboration. "Molecular imaging is really pushing the envelope of detection," he says. **L.S.**

sequencing technologies," he says, "If you have many markers of somewhat weak utility, the product of the probabilities of all those markers can be very powerful."

But despite all the effort going into biomarker discovery, very few have so far been turned into diagnostic tests. Hartwell says this is because validating them is a slow process. A year from now, he expects that the consortium of seven US research institutions that he leads in the search for breast-cancer biomarkers will have assembled 100 candidates. But they will then have to develop sensitive tests for those markers and only if the tests perform well in distinguishing diseased from healthy individuals, can they apply them to larger populations to measure their sensitivity and specificity.

Sudhir Srivastava, who heads the NCI's Cancer Biomarkers Research Group, says another factor delaying test development is that for the past 30 years, biomarker discovery has been driven largely by scientists doing basic research; such researchers may be less aware of which markers are likely to be most useful clinically. To try to change that, the NCI has brought together dozens of cancer-research institutions into an Early Detection Research Network (EDRN), which already funds a lot of research on detection, including Hanash's lung-cancer work. Bringing biomarkers to clinical application is one of the EDRN's main objectives. "This programme is attempting to strike a balance between discovery and validation," he says. "It brings clinical and basic scientists together, so that discovery is clinically driven."

The speed with which early-detection tests reach the clinic will also depend partly on how much of their development is taken up by pharmaceutical companies. Some fear that, because single biomarkers will turn out to be of limited value, drug firms will steer clear of detection and stick to treatment. Hartwell

disagrees. "Uniformly now, all the pharmaceutical companies want biomarkers for their particular drug, to know whether the drug is effective or not," he says. "The other place that this whole field could move into is drugs for prevention." As patients identified as high risk are likely to be taking preventive drugs for a lifetime, this could be a lucrative area for the drug industry.

From bench to bedside

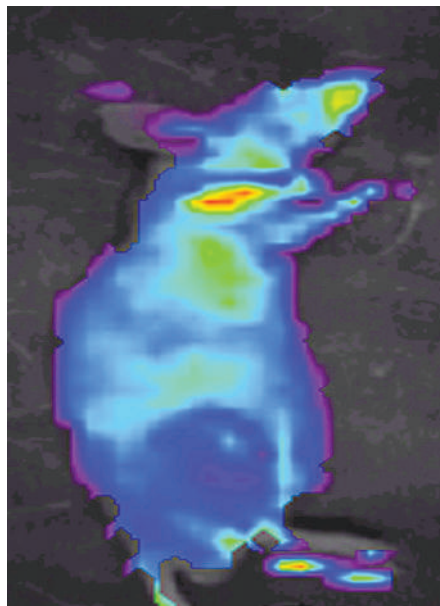
Drug companies are acutely aware of the potential in this area, agrees Brian Holloway, director of preclinical studies on Iressa (the drug that's used to treat lung cancer), at AstraZeneca in Macclesfield, UK. They are also keen to find ways of targeting their treatments more effectively, he adds. "There is no cancer drug that works for everybody, so drug companies will

look more and more for biomarkers to select the populations that will benefit from a given drug," he says.

As promising as early detection sounds, it does have its problems. For one thing, simply being able to detect the presence of a tumour may not give doctors enough information to treat a patient appropriately. Michael Pollak, director of the Program of Cancer Prevention at McGill University in Montreal, Canada, points out that at least half the world's people older than 60 have cancerous growths, most of which are non-aggressive and will not cause disease. "Early detection of all the lesions would result in a huge industry of medical interventions, most of which would be unnecessary, but a few of which would be life-saving," he says. "Early detection must be able to distinguish incidental from dangerous cancers — a much more difficult job."

And healthcare providers will still have to be convinced that early-detection tests are worthwhile at the population level, something that Richard Sullivan, head of clinical programmes at Cancer Research UK, says is notoriously hard to prove. It involves calculating the cost-effectiveness of the screening test, which in turn requires detailed information on the cost of screening, the long-term costs of cancer care, the sensitivity and specificity of the screening test and its impact on survival. But following Dr Papanicolaou's example, more screening programmes will almost certainly be established in years to come. ■

Laura Spinney is a science writer based in London and Paris.



The imaging techniques used to show up tumours (top red spot) are getting increasingly sensitive.

1. Gao, W. M. *et al.* *BMC Cancer* **5**, 110 (2005).
2. Jerónimo, C. *et al.* *J. Natl. Cancer Inst.* **93**, 1747-1752 (2001).
3. Pollak, M. N. & Foulkes, W. D. *Nature Rev. Cancer* **3**, 297-303 (2003).
4. Dinulescu, D. M. *et al.* *Nature Med.* **11**, 63-70 (2005).
5. Diehl, F. *et al.* *Proc. Natl Acad. Sci. USA* **102**, 16368-16373 (2005).

Off by a whisker

Much of what we know about cancer comes from studying mice, and potential therapies are tested in the animals. But the differences between the species can scupper the best laid plans of researchers and drug companies, reports **Carina Dennis**.

It was in 1991 that Bob Weinberg first realized he had a problem with mice. He and his postdoc Tyler Jacks were trying to develop a mouse model for retinoblastoma, a childhood cancer of the retina. It results from the loss of a gene called *Rb*, so the team genetically engineered mice to lack the same gene. But the mice didn't get retinoblastoma. Instead, they developed tumours in their pituitary glands. The finding shocked Weinberg. "Up until then, I had always believed that all mammals were biologically equivalent," he says. "This planted the seeds of doubt in my mind."

Weinberg, based at the Whitehead Institute for Biomedical Research in Cambridge, Massachusetts, is one of the pioneers of the molecular age of cancer research. He was involved in the early work on the first human cancer-causing and cancer-suppressing genes in the early 1980s. But when he saw that mutations in such genes didn't cause the same kind of cancer in mice and humans, he began to ask himself why. He became aware of other examples that challenged researchers' faith in how accurately mice could replicate human tumours, and has since sought to bring this to his colleagues' attention¹. "There is a laundry list of problems with mouse models of cancer," he says.

Although some genetically engineered mice have come a long way since the early days, many researchers concede that other kinds of mouse model remain especially problematic. Of the potential anticancer drugs that give promising results in tests on mice with cancer, only about 11% are ever approved for use in people². It is also possible that drugs that would have worked in people failed in pre-clinical mouse trials, although there is no way of knowing. "I'm not a naysayer who thinks we should discard valuable mouse models, thereby throwing the baby out with the bathwater," says Weinberg. "My rallying cry is that we need to be working on this problem." Now many researchers, Weinberg included, are doing just that, by making mouse models more faithful replicas of human cancer development.

Different strokes

Mice are still crucial for cancer research. After all, there is only so much one can learn from studying cells in a lab dish. Cancer is a complex, three-dimensional disease that changes, evolves and spreads through the body. Mice, easy to breed and genetically

manipulate, and with a completed genome sequence, are the obvious choice for scientists wanting to pick apart these processes and test new drugs. And, as mammals, they share enough biology with us to make some useful comparisons possible.

But despite having broad similarities, mice have significant differences that can scupper cancer experiments (see graphic, overleaf). For one thing, most mouse tumours originate in different types of tissues from ours and, unlike us, their healthy cells can maintain the ends of their chromosomes, a key factor influencing which mutations tumour cells develop. One of the key mouse models, especially in drug testing, involves grafting human

cancer cells into mice and seeing how the resulting tumours respond to treatment. But human cells are likely to behave differently in a mouse than in a human body, making results hard to interpret.

Some of these problems are insoluble, arising as a result of fundamental differences between the two animals. But others can be fixed. In recent years, researchers have developed

tricks to genetically engineer mice, or tinker with the human cells they graft into them, that more accurately reflect cancer in humans. And now both academia and industry are keen to put these mice to the test and see whether they can better predict how humans will respond to new drug compounds. "It took a while to make mouse models that accurately reflect human diseases but now that we have them, now is the time to use them," says David Tuveson of the University of Pennsylvania in Philadelphia.

Early attempts to genetically engineer mice to develop cancer were, in retrospect, simplistic: turning a cancer-associated gene on or off in every cell in every tissue of the mouse.

These mice often developed tumours, but they rarely looked anything like the human cancer, as Weinberg discovered. What's more, the technology used to alter mouse genes took effect early in development, whereas most human tumours develop later in adult life, following a stepwise series of mutations that turn a normal cell into an invasive cancer.

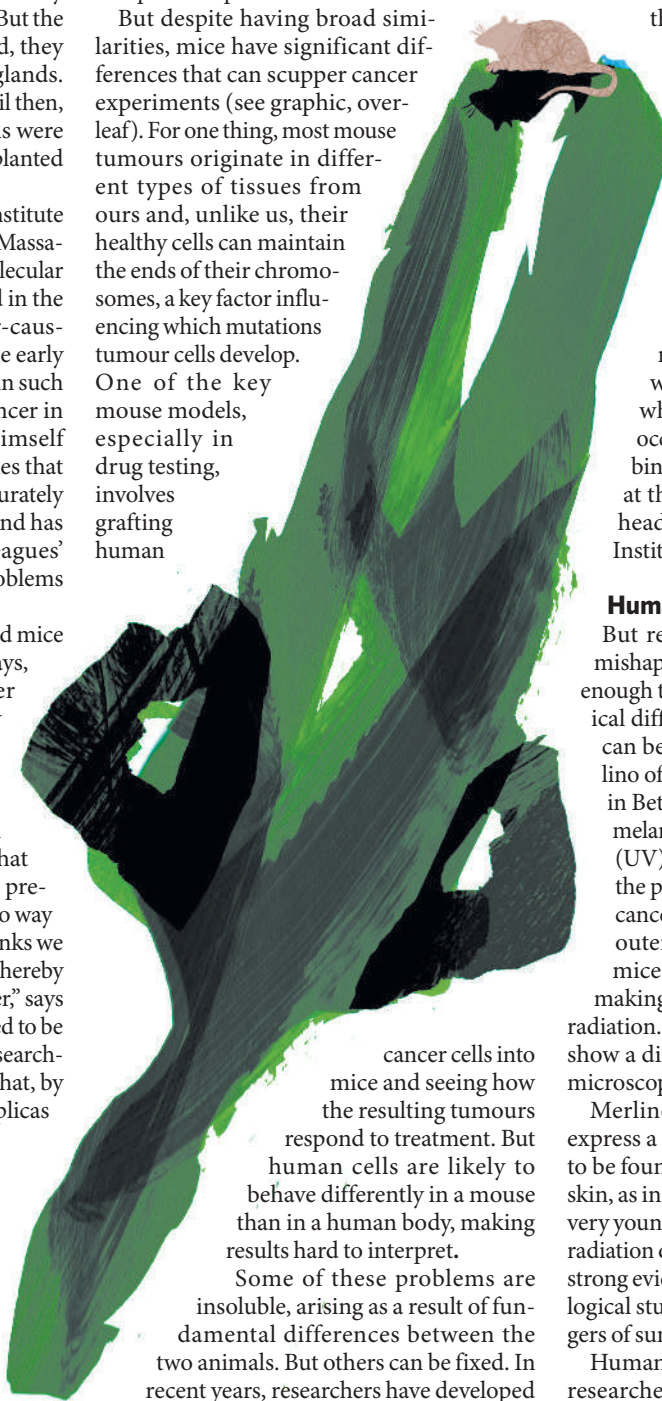
Over the past decade, researchers have been able to recreate these steps more accurately. A key breakthrough was being able to control when and where the cancer-causing mutations occurred³. "We're getting the right combination of mutations at the right place at the right time," says Jacks, who now heads his own lab at the Massachusetts Institute of Technology in Cambridge.

Human touch

But recreating the sequence of genetic mishaps in a cluster of cells isn't necessarily enough to replicate human cancer. The physical differences between humans and mice can be a major obstacle, says Glenn Merlino of the National Cancer Institute (NCI) in Bethesda, Maryland. Merlino works on melanoma, which is induced by ultraviolet (UV) irradiation of the skin. Melanocytes, the pigment-producing cells that become cancerous in melanoma, are found in the outermost layer of human skin, but in mice, they are confined to hair follicles, making it harder to study the effects of UV radiation. What's more, the resulting tumours show a different cellular structure under the microscope.

Merlino's team engineered mice to over-express a cell signal that causes melanocytes to be found in the outer layers of the rodents' skin, as in humans. They went on to show that very young, but not adult mice, exposed to UV radiation develop malignant melanoma, giving strong evidence to back up previous epidemiological studies highlighting the potential dangers of sun exposure in children⁴.

Humanized mice such as these could help researchers tackle some fundamental ques-



S. PEMBERTON

Vive la difference

While some scientists are bent on making the mouse more human-like, others are celebrating our differences. Comparing mouse cancers with human ones helps researchers cut through the complexity of cancer and discover new genes and processes key to the disease. "It can act as an evolutionary filter to sift through the noise," says Jeff Green, a molecular biologist at the National Cancer Institute in Bethesda, Maryland.

Comparing gene activity in mouse and human tumours can help reveal the biochemical pathways involved in the development of cancer that would otherwise be obscured by the complexity of the data obtained from humans alone. Green, for example, is using

this method to uncover the biochemical pathways involved in oestrogen signalling that are present in both species.

The idea is that pathways that have been conserved through the millions of years of evolution that separate humans and mice are likely to play a crucial role in the basic cell biology of breast cells, and so might play a role in cancer.

Tyler Jacks, of the Massachusetts Institute of Technology in Cambridge, on the other hand, used this strategy to uncover new mutations and altered signalling in human lung cancer. The work revealed that a gene called *KRAS2* is often mutated — and the signalling pathway of which it is part perturbed — in a subset of lung cancers⁹.

Seeing how the human and mouse genomes compare is also yielding new clues. Lynda Chin of the Dana-Farber Cancer Institute in Boston, Massachusetts, and her colleagues recently used a method that highlights differences between the two genomes to identify a gene that enhances the spread, or metastasis, of the skin cancer melanoma¹⁰.

The team found that one region of the mouse genome had extra copies in those tumour cells with the potential to metastasize. The equivalent region in human melanoma cells also turned out to have extra copies. The team subsequently pinpointed a gene in the region called *NEDD9*, which was overactive in tumours and made cells more invasive.

Meanwhile, Scott Lowe of the Cold Spring Harbor Laboratory, New York, and his colleagues used a similar strategy to compare the genomes of human tumours and that of a mouse model of liver cancer.

They identified a region of the mouse genome that has extra copies present, found similar copies in human tumours, and showed that the overactivity of two genes within this copied region accelerates tumour growth¹¹.

Both Chin and Lowe's work show how the mouse can act as a filter to find new genes involved in cancer formation, maintenance and metastasis — and spotlighting potential new drug targets. "You can glean more information from looking at both animals, than either in isolation," says Jacks. **C.D.**

tions — such as where and how cancer begins. "The problem is we don't know much about the cell-of-origin in human cancers," says Jacks. Some suspect that it is the stem cells that repair and renew adult tissues (see 'The root of the problem', page 742). "Those are hard questions to get to in humans. This is an example of the value of mouse models — they are highly manipulable, for which there is no comparison in human," says Jacks.

The latest generation of mouse models mimics the human disease process more faithfully. But, in therapeutic terms, the real test is how well they reflect the human response to drugs. Early signs are encouraging⁵. "New mouse models hold a lot of promise. But now they have to show what they are worth," says Anton Berns, of the Netherlands Cancer Institute in Amsterdam.

Foreign bodies

The pharmaceutical industry has typically used 'xenograft' models to screen new drugs. In these, human cancer cells are injected under the skin of a mouse with a deficient immune system, to prevent rejection of the tumour. Although virtually every successful cancer drug on the market will have undergone xenograft testing, many more that show positive results in mice have had little or no effect on humans, possibly because the human tumours are growing in a foreign environment. "In many cases, the mouse is just a container for the human cells," says Roberto Weinmann, the director of oncology at pharmaceutical company Bristol-Myers Squibb in Princeton, New Jersey.

Many companies therefore are interested in the new genetically engineered mouse models. "We are in a transition," says Giulio Draetta, head of basic cancer research at drug company Merck, based in Boston. "We are now build-

ing a substantial operation using transgenic models and are investing in licensing genetic models that currently exist as well as building our own models," he says.

"We still rely very heavily on xenografts, as do most companies. But, as the models get better, we are watching the field very carefully," adds Richard Gaynor, vice-president of cancer research at Eli Lilly in Indianapolis.

Some argue that a major factor preventing industry fully embracing transgenic mice are the patents associated with certain kinds of mice engineered to develop cancer, in particular the OncoMouse patents owned by chemical company DuPont, which constrains the commercial development of such mice⁵. Some think the patents have deterred drug companies from embracing transgenic mouse models. The licensing arrangements that commercial groups need to negotiate with DuPont have made using such models "more arduous and expensive", says Gaynor, although academics and non-profit groups can use the technology freely.

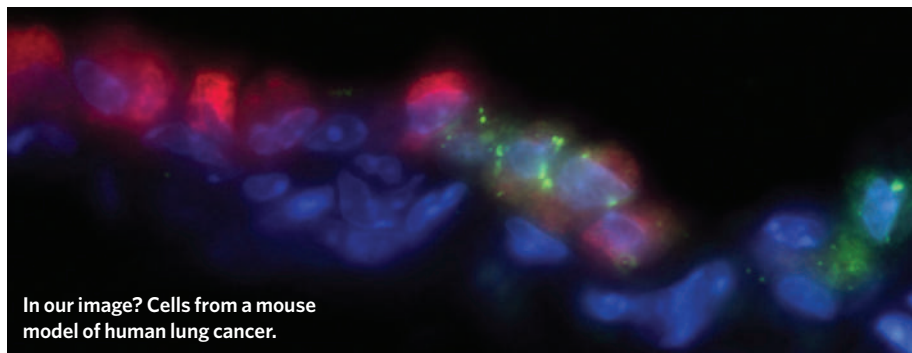
Intellectual-property issues aside, xenografts will remain many companies' model of choice for the foreseeable future. They can be ready within weeks, whereas transgenic mouse mod-

els can take months to develop tumours and are often unpredictable in where and when they develop them. As hundreds of mice are required for screening, xenografts are quicker and cheaper.

Making connections

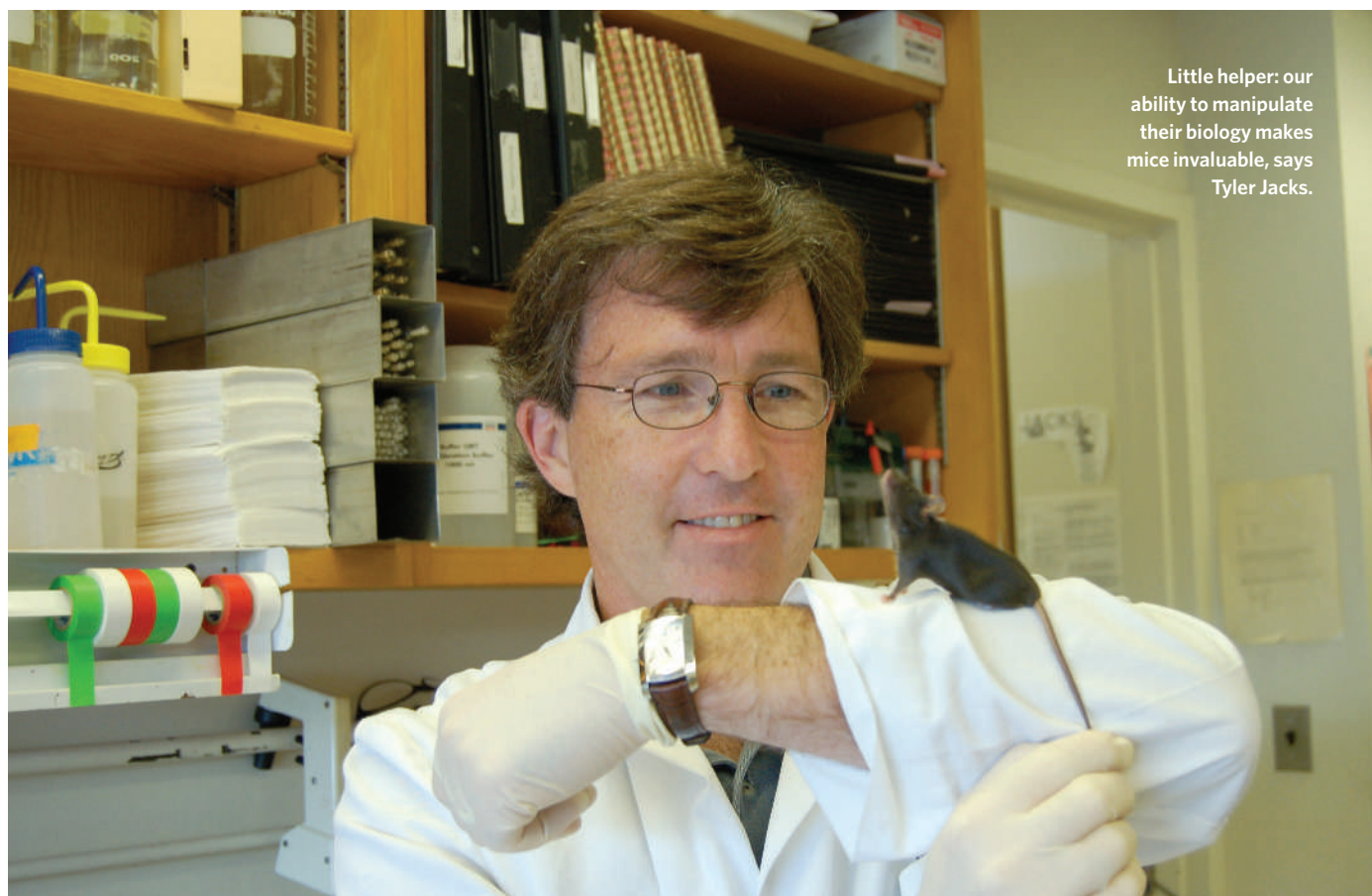
One possible way to improve xenografts is to manipulate the cellular environment into which cancer cells are injected. Connective tissue cells, or stromal cells, play an important role in the development of a tumour, often supplying it with the cell signals and nutrients it needs. But interactions between tumour cells and their neighbours are often lost in xenografts, because proteins from one species can't interact with their counterparts in the host. "The consequence is that the tumour often looks nothing like that seen in the patients," says Weinberg.

Weinberg is among a number of scientists who have addressed this problem by putting human connective tissue cells into mice along with the cancer cells. The resulting tumours have a more human-like structure and metastasize more like human cancers do^{6,7}. "It gets us closest to the human situation," says Ronald DePinho of the Dana-Farber Cancer Institute



In our image? Cells from a mouse model of human lung cancer.

KIM ET AL. CELL 121, 823-835 (2005) ELSEVIER



Little helper: our ability to manipulate their biology makes mice invaluable, says Tyler Jacks.

A. DECONINCK

in Boston, Massachusetts, who has co-founded the company AVEO Pharmaceuticals to develop tissue-specific cancer models, including breast cancer.

Adding human cells into the mix is one approach. Another way to make mice respond to drugs in a more human-like fashion is to give the mouse human genes. The driving force behind this is the considerable differences in the way that mice and humans metabolize drugs. For instance, two human enzymes CYP2D6 and CYP3A4, which together metabolize more than 70% of drugs on the market, have markedly different activities compared with their rodent equivalents⁸. The consequence is that mouse models may be of limited value in predicting the effectiveness or toxicity of drugs in humans.

Mine of information

Researchers have tackled this problem by engineering genes encoding some of the human enzymes into the mouse⁸. Studies of these 'humanized' mice indicate they can process some drugs in a more human-like manner. Combining these humanized mice with revamped mouse cancer models could be a powerful way to test drug candidates.

In the future, researchers may broaden the types of human genes they add in to the mouse. "My hope is that we will be able to create mouse cells and tissues that may closely resemble human tissues," he says.

But there is a limit to how far scientists can

get mice to be like humans. So other researchers are mining human clinical data for insights into the cancer process, then using this to inform mouse studies. In the past, clinical data have often been collected on an ad hoc basis. But now researchers are tackling the process in a systematic and comprehensive fashion. Assisted by the Internet and better bioinformatics tools, there are large-scale efforts to collect and catalogue data on a wide range of human tumours.

Howard Fine of the NCI is leading one such initiative. The aim is to perform genetic and molecular analyses of samples of human brain tumours sent to the NCI. Data are correlated with each patient's clinical course. The study is a pilot for the NCI's larger cancer Biomedical

Informatics Grid, which will provide a global network for researchers to input information and access bioinformatics tools for mining cancer data. The study, which currently has data for 700 tumours and will eventually contain information for 2,000 tumours, is already yielding results, says Fine, who hopes to publish the findings shortly.

Researchers are increasingly shuttling between human and mouse, using information from human data to refine mouse models and using insight from the mouse to uncover new disease processes and test predictions for clinical responses. But industry and the scientific community shouldn't expect too much from the humble rodent. "Mice are valuable but they are, after all, still mice," says Fine. "The best study subject will always be the human."

Carina Dennis is Nature's Australasian correspondent.



HOW MICE DIFFER FROM HUMANS

- Cancers tend to form in different types of tissue
- Tumours have fewer chromosomal abnormalities
- Ends of chromosomes (telomeres) are longer
- Telomere-repairing enzyme (telomerase) active in cells
- Short lifespan
- Fewer cell divisions (10^{11}) during life than humans (10^{16})
- Metabolic rate seven times higher than humans
- Lab mice highly inbred and genetically similar

1. Rangarajan, A. & Weinberg, R. A. *Nature Rev. Cancer* **3**, 952-959 (2003).
2. Kola, I. & Landis, J. *Nature Rev. Drug Discov.* **3**, 711-715 (2004).
3. Johnson, L. et al. *Nature* **410**, 1111-1116 (2001).
4. Noonan, F. P. et al. *Nature* **413**, 271-272 (2001).
5. Sharpless, N. E. & DePinho, R. A. *Nature Rev. Drug Disc.* doi:10.1038/nrd2110 (2006).
6. Chudnovsky, Y., Adams A. E., Robbins P. B., Lin Q. & Khavari P. A. *Nature Genet.* **37**, 745-749 (2005).
7. Kuperwasser C. et al. *Proc. Natl Acad. Sci. USA* **101**, 4966-4971 (2004).
8. Gonzalez, F. J. & Yu, A.-M. *Annu. Rev. Pharmacol. Toxicol.* **46**, 41-46 (2006).
9. Sweet-Cordero, A. et al. *Nature Genet.* **37**, 48-55 (2005).
10. Kim, M. et al. *Cell* **125**, 1269-1281 (2006).
11. Zender, L. et al. *Cell* **125**, 1253-1267 (2006).

The root of the problem

Is targeting cancer stem cells a way to finish tumours off once and for all — or just the latest in a long line of false dawns? **Alison Abbott** looks at a debate that's generating both heat and light.

"**Y**our cancer has returned." The words that patients most dread are the ones that most of them eventually hear. Why tumours that had been shrunk to nothing by a first round of treatment should recur has long been a mystery. Now, a growing number of researchers think it may be because biologists have overlooked a key part of how cancers grow.

It's all down to stem cells, they say. Normal stem cells repair and renew our tissues. Recently, research has focused on the idea that a kind of evil twin, a cancer stem cell, could stock a tumour with cancer cells, rather like the way a weed can grow back from its roots. If the idea is right, then cancers return because existing treatments cut down the leaves and shoots but leave the roots intact.

According to the theory, cancer stem cells have a physiology that makes them particularly resistant to typical treatments. They lurk for months or years, then re-seed the cancer, usually in a more aggressive, often fatal, form (see 'Why the drugs don't work'). Find a treatment that targets the cancer stem cell, without harming normal stem cells — as researchers are already trying to do — and scientists might be able to protect against recurrence.

It's a controversial idea. The theory is promising, but the field is awash with hype. Researchers observe wryly that more reviews seem to have been published than primary papers. Ideas abound about how the theory explains every aspect of cancer, and about where and how to find therapeutic targets. Hard data are emerging only slowly.

Sean Morrison, a stem-cell biologist at the University of Michigan, Ann Arbor, who reported one of the first key differences between normal and cancer stem cells in May¹ remains downbeat. "Remember apoptosis, remember angiogenesis?" he asks rhetorically. These properties of cancers — respectively, their ability to bypass the normal cell-suicide response to damage, and their ability to create their own blood supply — were once seen as routes to powerful therapies, but have yet to live up to their initial promise.

Yet whatever their clinical potential, the idea of cancer stem cells is prompting biologists to think about cancer in a new way.

Many different tissues harbour adult stem cells. The cells don't normally divide much, but when prompted by death in their realms, they

do so asymmetrically.

One of the two daughter cells is a replica of the parent cell — the property of self-renewal — and the other is a 'progenitor' cell that goes on to differentiate into more specialized cells. Neural stem cells, for example, can replace both of the brain's major cell types, neurons and glia. Blood stem cells can replace myriad red and white blood cells.

Cancer stem cells share the ability to self-renew and to spawn many different kinds of cell. They were identified in leukaemias in the late 1990s. John Dick at the University of Toronto harvested leukaemic cells from patients and found that only a small proportion could reproduce the same cancer in immune-deficient mice². The ability to recapitulate the same cancer is what defines a cancer stem cell.

In the blood

Leukaemia is relatively easy to work with. You can access the cancer cells by drawing blood, and blood stem cells are well-characterized, particularly in terms of the presence or absence of specific marker proteins that define 'stemness'. Getting hold of fresh a solid tumour is more difficult, and much less is known about the relevant markers. That said, cancer stem cells have now also been demonstrated in breast³ and brain⁴ tumours. They bear convincing markers of stemness, and they can recapitulate the cancer.

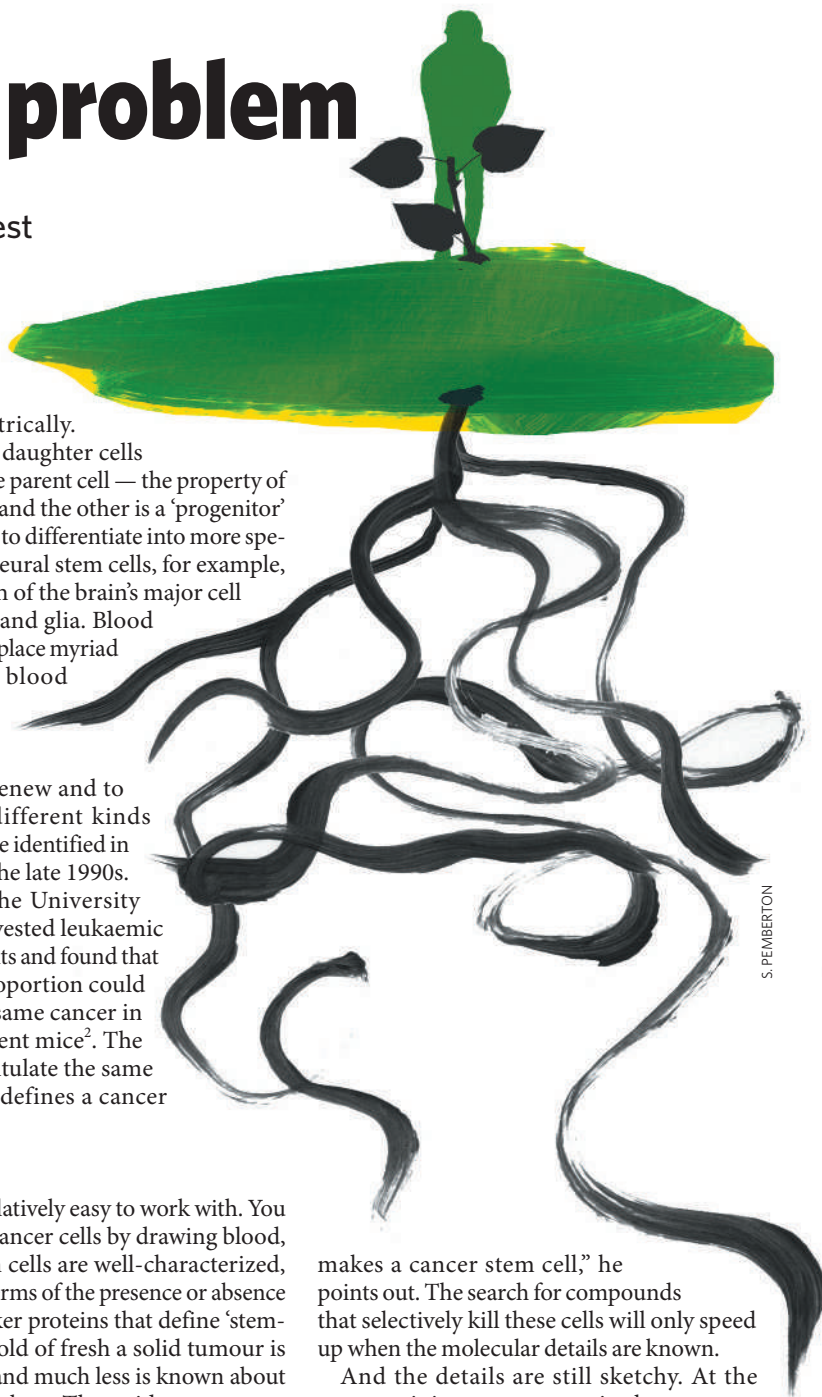
There are also many early reports of stem-like cancer cells in other solid tumours such as ovarian, lung and skin cancers that have yet to meet all the criteria of proof. "There is no question in my mind that cancer stem cells will prove to be the rule rather than the exception," says Michael Clarke, of the Stanford Comprehensive Cancer Center. It's just a question of time to get the results, he thinks.

On the other hand, oncologist William Hahn of Harvard University — who describes himself as "a sceptical observer who likes the idea of cancer stem cells" — is more circumspect. "There is no real molecular definition of what

makes a cancer stem cell," he points out. The search for compounds that selectively kill these cells will only speed up when the molecular details are known.

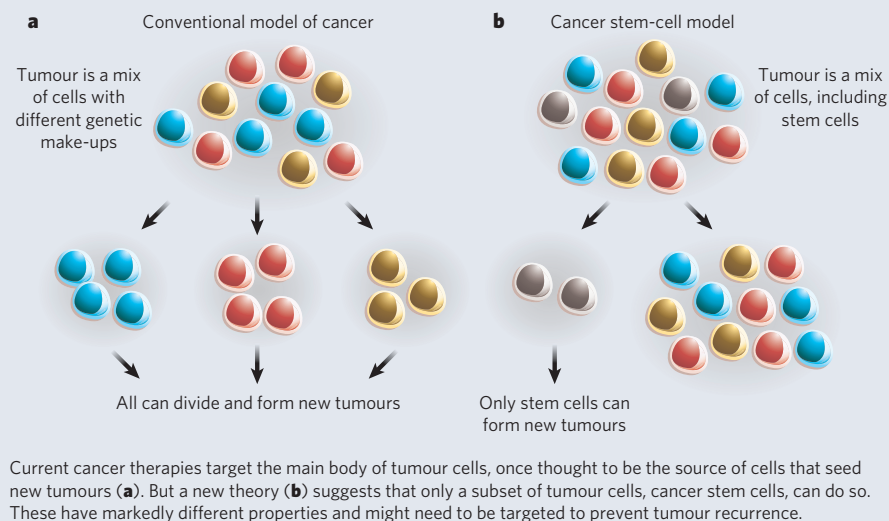
And the details are still sketchy. At the moment it is not even certain that a cancer stem cell derives directly from a normal stem cell, although this will probably often be the case. Cells become cancerous when they have accrued a critical number of mutations in genes in the interacting signalling pathways that control growth, allowing the cells to evade the normal mechanisms that restrict growth. Unlike many cells, stem cells live long enough to acquire the required changes.

But researchers have shown that a non-stem cell that picks up a mutation giving it the ability to self-renew might also develop into a cancer stem cell⁵. Different cancers probably contain cancer stem cells of different origin. But the origin of cancer-initiating cells is less important to those seeking therapeutic targets than finding molecules in them that they do not share with tissue stem cells.



S. PEMBERTON

WHY THE DRUGS DON'T WORK



between stem cells that generate tumours and those that do not," says Morrison.

Also getting a cancer stem-cell spin are the 'drug pumps'. These membrane proteins, which have long been therapeutic targets, remove drugs and other foreign molecules from a cell. They are believed to be a major cause of the resistance that cancer cells develop to drugs that initially shrink tumours.

Resistance cell

Stem cells are assumed to be built like fortresses, with multiple mechanisms to protect them from toxic environmental agents, so they don't replace tissue with damaged goods. Cancer stem cells might also be packed with drug pumps, explaining their presumed extreme resistance to therapies. Some suspect the unexpected occurrence of resistance to Gleevec — the first anticancer agent targeted to a specific signalling-pathway component, called BCR-ABL — could be down to cancer stem cells' ability to pump the drug out. Dean, for example, is screening compounds in the hope of targeting such drug pumps in cancer stem cells.

A different approach to hitting cancer stem cells is to make highly specific antibodies that can seek out and kill them. This requires a deep understanding of the proteins found only in cancer stem cells, which antibodies could target. It seems a tall order, but a number of

biotech companies are trying to develop such methods.

Outside the search for therapeutic targets, established themes of cancer are also being recast in the light of the stem-cell concept. Could it for example, give insights into metastasis, given that normal stem cells can relocate to repair distant tissue? Could it explain why chronic injury frequently results in cancer, as in such circumstances stem cells are called upon to continuously replace tissue and thus increase their chances of acquiring mutations?

Even those scientists who fear that the cancer stem cell may prove to be a sidetrack don't believe it will be a dead-end — simply because such a lot of cancer biology is being reconsidered, and new things uncovered, en route. And if the concept performs beyond expectation, then the root-killing generation of therapeutics will emerge to ensure permanent remissions, to protect patients from deadly cancer recurrence.

Alison Abbott is Nature's senior European correspondent

Given that so little is known about marker molecules in any sort of stem cell, at the moment it's a bit like trying to do a crossword without all the clues. But a couple of groups have identified two molecular signalling pathways that are regulated differently in the two stem-cell types. And these present the first leads for potential therapies targeting cancer stem cells.

Flower power

Craig Jordan, a molecular biologist at the University of Rochester, New York, has identified a molecular signalling system that is permanently switched on in leukaemia stem cells but not in blood stem cells. He found that the molecule parthenolide, the main active ingredient in feverfew, a herbal medicine used to treat migraines, blocks a key component of this system. This prompts leukaemic stem cells to commit suicide, whereas normal stem cells are unaffected⁶. Jordan says he hopes to have a derivative of the molecule in clinical trials by the end of the year.

Morrison's discovery was that deleting a gene called *Pten*, commonly inactivated in human leukaemia, stimulates the growth of leukaemia stem cells but inhibits the growth of normal blood stem cells. And he found that the drug rapamycin reversed the effect of the gene deletion in mice, leaving the cancer stem cells unstimulated, and reactivating the growth of blood stem cells¹. Rapamycin is used clinically as an immunosuppressant to prevent transplanted organs being rejected, and several variants happen to be in clinical trial for cancers. His group's findings are sure to affect how the trial results are interpreted, says Morrison.

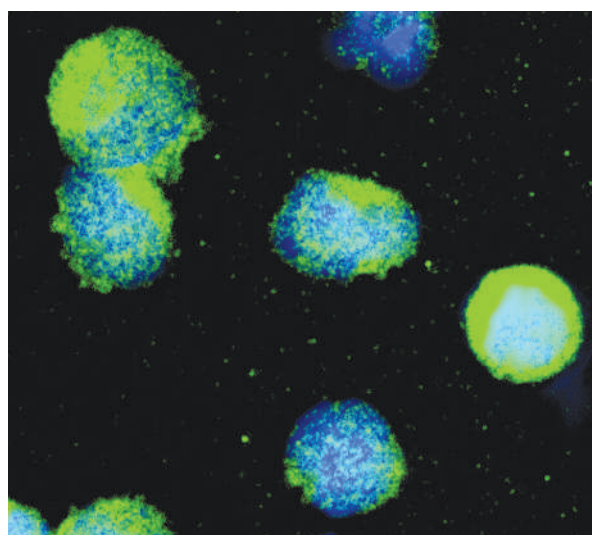
Other pathways that are often aberrantly activated in cancer may or may not be differently regulated in normal

and cancer cells. These pathways have been pillars of mainstream cancer research for years, and are now acquiring a stem-cell spin. Two of the many contenders are the *Wnt* and *Sonic hedgehog* pathways which, in their healthy manifestations, play a key role in embryonic development, stem-cell renewal and tissue repair.

"There is no proof that *Sonic hedgehog* is involved in cancer stem cells but my intuition is that it will be," says molecular biologist Michael Dean of the National Cancer Institute in Frederick, Maryland. He is already looking for candidate drugs, initially using cyclopamine, a molecule found in the poisonous lily *Veratrum californicum*. Cyclopamine interrupts *Sonic hedgehog* signalling in prostate, gut and brain cancers.

The hard core of cancer-stem-cell researchers view this tendency to hang a new collar on an old dog with caution. "There are a lot of hypotheses, but you have first to distinguish

"Some of the pillars of mainstream cancer research are getting a stem-cell spin."



Leukaemia stem cells have the ability to seed new cancers.

1. Yilmaz, O. H. et al. *Nature* **441**, 475–482 (2006).
2. Bonnet, D. & Dick, J. E. *Nature Med.* **3**, 730–737 (1997).
3. Al-Hajj, M. et al. *Proc. Natl Acad. Sci. USA* **100**, 3983–3988 (2003).
4. Singh, S. K. et al. *Nature* **432**, 396–401 (2004).
5. Krivtsov, A. V. et al. *Nature* **442**, 818–822 (2006).
6. Guzman, M. L. et al. *Blood* **105**, 4163–4169 (2005).

Conservation requires multiple approaches

SIR — In your News Feature “The tiger’s retreat” (*Nature* **441**, 927–930; 2006) you highlight the differing approaches to the protection of India’s rapidly diminishing tigers and biodiversity taken by the Wildlife Conservation Society and the Ashoka Trust for Research in Ecology and the Environment (ATREE).

Although strategies might differ, our ultimate goal is the same: to save as much biodiversity (including tigers) as possible.

In some areas, relocation of people out of protected areas, as advocated by Ullas Karanth of the Wildlife Conservation Society, may be the only option. In others, local communities can be effective allies and partners in conservation, as suggested by C. Madhegowda and Nitin Rai of ATREE. A heterogeneous world — and particularly a country such as India — needs multiple complementary approaches (K. S. Bawa, P. H. Raven and R. Seidler *Conserv. Biol.* **18**, 859–860; 2004).

The threats to Indian biodiversity are varied, from invasive species and poaching to infrastructure development. The focus on local communities as major drivers of change, when they are often themselves victims of these pressures, precludes a consideration of the multiple factors degrading protected areas. And the focus on tigers runs the risk of pushing other important components of biodiversity conservation into the background. In addition to the external pressures, we have to worry about governance and policies.

Archaic governance can produce protected areas with no people but little biodiversity. Conversely, good governance can allow biodiversity in protected areas to flourish even with people in or around them. Integrated approaches to conservation, involving the participation of local people, will thus remain indispensable into the foreseeable future.

Kamaljit S. Bawa

Department of Biology, University of Massachusetts, Boston, Massachusetts 02125, USA, and Ashoka Trust for Research in Ecology and the Environment, 659 5th A Main Road, Hebbal, Bangalore 560 024, India

Researchers should ensure that their actions are lawful

SIR — Your News story “From aircraft engineer to FBI suspect” (*Nature* **442**, 232; 2006) highlights academics’ naïveté about security issues.

In my opinion, the motives of researchers do matter. Unfortunately, most media portray scientists concerned in these matters, when

they come to light, as innocent victims rather than as people who need to be more aware of the risks and who should ensure that their actions are lawful.

In the case you describe in your News story, it is incredible that J. Reece Roth should not have realized that hiring a Chinese graduate student might pose potential security risks. Anyone signing an agreement to receive military government contracts and funding is required by law to sign a waiver stating that the research will not be shared with foreign powers and that all involved must be US citizens. This is standard policy in military-awarded contracts with secret and security-related applications.

Peter Cohen

5507 Englishman Place, Rockville, Maryland 20852, USA

Wiki and other ways to share learning online

SIR — As described in your News story “Online methods share insider tricks” (*Nature* **441**, 678; 2006), the ability to share detailed technical advances and modifications of existing laboratory protocols is a benefit of web-based publishing. However, there are potential problems with using a wiki-style mechanism for publication of practical procedures. If an original protocol can be edited by any user, without a checking and editing mechanism, then there is the potential for website owners to be seen to promote incorrect and potentially hazardous procedures.

The OpenWetWare site described in your News story is an excellent example of a more ‘managed’ wiki-style site, but there are other ways of providing dynamic, interactive technical information online. An example that has been operating for several years in the chemistry community is Synthetic pages LLP (www.syntheticpages.org), on whose behalf I am writing. This provides a forum for the dissemination of practical synthetic chemistry procedures online in a different format to a wiki.

Users are encouraged to submit details such as reproducibility, scale of reaction, unsuccessful attempts and modifications to procedures. The submissions are then reviewed by the editors using a ‘light-touch’ process. Ultimately, more than 95% are published. Original submissions cannot be directly edited by users, but attributed comments and modifications are encouraged and added at the end of the page. Registered users can hyperlink between pages using the ‘comments’ facility or within their submission of a new page, so people can see the protocol evolve and see which users have been involved in this process, enabling decisions on how best to use the information.

All the pages on the site can be browsed or searched free of charge and without any form of registration, but users who wish to contribute must register. The site contains pages from all over the world. It has more than 3,600 users and more than 10,000 unique search visits per month.

Although different from a wiki, the format offers a dynamic, interactive style of web-based publishing which could be more valuable to the community because of its managed style, and because the editorial approach enhances the safety and overall value of a submission and subsequent modifications or comments. It is a format that could be extended to many other areas of experimental research, such as biology, engineering, medicine and others.

Stephen Caddick

Department of Chemistry, University College London, 20 Gordon Street, London WC1H 0AJ, UK

It’s easier to patent plants than to publish research

SIR — Your timely News story “Is India’s ‘patent factory’ squandering funds?” (*Nature* **442**, 120; 2006) could open a Pandora’s box about the prevailing state of affairs in academic research institutions in the developing world. Obtaining a US patent, especially on plant varieties, is an easy alternative to publishing in peer-reviewed journals for high-profile scientists, because money for filing patents is easily available, with no questions asked about the financial viability of the discovery.

The bare truth is that one can pick any vegetatively propagated plant where no prior patent exists, define its identification characteristics and potential benefits for cultivation, pay a fee to the US Patent and Trademark Office and obtain a patent number for subsequent professional rewards — with no demonstrable benefit to biology.

One of the examples cited in your News story, a cow-urine distillate to enhance the activity of antibiotics (US patent 6410059), has been criticized for not having been through scientific scrutiny (see, for example, a report in *The Hindu* newspaper at www.hinduonnet.com/seta/2002/09/19/stories/2002091900150300.htm). The claim has never been substantiated by peer-reviewed publications.

This kind of activity, which is widespread, diverts millions of dollars from research into filing patents. The publicly funded scientific institutions concerned should ignore such CV-enhancing practices in favour of published research.

U. C. Lavania

Central Institute of Medicinal and Aromatic Plants, Lucknow 226 015, India

BOOKS & ARTS

Defining moments

What are the major principles that controlled the origin of life?

Singularities: Landmarks on the Pathways of Life

by Christian de Duve

Cambridge University Press: 2006. 274 pp.
£30, \$48

David Penny

Christian de Duve seeks to understand the mechanisms underlying the major changes that occur during the origin and history of living systems — the singularities. These include the origin of the protometabolism, membranes, protein, DNA and the prokaryote–eukaryote divide. He wants to know whether such transitions are necessary, almost inevitable, or very unlikely (even though they happened anyway). Are there alternatives that were essentially as good, or potentially better? Were the choices externally or internally imposed? And so on.

Singularities — the latest book to be tagged as de Duve's last — is in the tradition of *The Major Transitions in Evolution* by John Maynard Smith and Eörs Szathmáry (W. H. Freeman, 1995), which covered the aforementioned transitions but focused more on later stages, such as meiosis, multicellularity and language. Most of de Duve's book, in contrast, centres on the origin and nature of life, although he does go on a thoughtful scamper from the last universal common ancestor, the origin of eukaryotes, and multicellularity, before reaching the arrival of humans. Most of the book, then, is the great quest for the origin of life, next to which Frodo and Sam's journey into Mordor to destroy the Ring pales into a simple tourist trip.

De Duve poses a very important question: what are the fundamental issues that have to be 'solved' to get life as we know it? In evolutionary biology over the past few decades, we have mainly heard the 'contingency' message, which emphasizes the stochastic nature of evolution and focuses on the details. As we have heard countless times, if we rerun the tape of life again we will get something a bit different, and so on. Certainly, we are unlikely to get a horse-like creature with serine at position 23 of a β -haemoglobin, and a human-like animal with alanine at the equivalent position. But focusing only on contingency misses the more fundamental questions of major principles to which life might be conforming, and whether some general outcomes are predictable. This is the approach taken by de Duve.

As an example, consider Jeremy Knowles'



G. SILK/TIME & LIFE/GETTY

Tougher than Mordor — the quest to understand the large changes that led to the development of life.

view that an enzyme has catalytic perfection if it is limited by the rate of diffusion of a small substrate molecule onto the enzyme; no change to the properties of the enzyme can speed up the rate of the reaction. Because archaea, bacteria and eukaryotes can each have an equivalent enzyme that is 'perfect', but with very different protein sequences, there must be millions of different sequences that will carry out the reaction at the maximal speed. Now isn't this a much more fundamental and interesting conclusion than just that the sequences differ?

We can either concentrate on the principles, or we can focus on the details. Yes, the details will differ if the tape of life is rerun, but will we get the same basic metabolism and energy sources? Would we start again with RNA? Will we get proteins as the primary catalysts? Will there be a mix of primary producers, herbivores, carnivores and saprophytes, as well as small bacterial-like cells and larger-celled predators? Why do we not have endothermic plants? These are the big, interesting questions. De Duve focuses on these principles, refusing to be sidetracked by contingencies.

The view that de Duve argues is far more darwinian than many have used over the past few decades. Charles Darwin learned from the early statistical interpretations of Adolphe Quetelet that we could make firm predictions

about random events if we had really large numbers. Even a rare event will be effectively certain to occur if there are millions of trials over millions of years. From this, you might gather that I like de Duve's contemplative approach. Yes, contingency and the stochastic nature of evolution are very important, but there is much more besides.

De Duve's general approach is fine, then, but how about the specifics? He favours a weak version of the RNA world in which most RNA catalysts are relatively short and are assisted by short (non-coded) peptides combined with other compounds that he calls multimers. It is unfortunate that we don't expect to find much evidence for such relics in modern metabolism — in contrast to cofactors (such as NADH and FAD), which are generally interpreted as relics of catalytic RNA from the RNA world. If there were such multimers, the amino-acid parts at least could be incorporated directly into the protein backbone of modern proteins, and therefore not be recognized as relics.

What are the next experiments? Are the chemical reactions that support life easier or harder at a pressure of hundreds or thousands of atmospheres? Are dilute or high concentrations advantageous? All life now occurs at high concentrations, restricted by membranes.

The biologists' top-down approach is making

good progress in simplifying living systems to their basics. However, chemists seem to be having a harder time with their bottom-up approach.

The mathematics of Elchanan Mossel and Mike Steel argue that autocatalytic cycles, given relatively simple assumptions, are expected to occur in simple non-living systems. Experiments with *in vitro* RNA evolution are able to test many trillions of sequences simultaneously. Could chemists carry out trillions of reactions simultaneously? Perhaps the new approach of evolving proteins in oil droplets

(*Nature* **440**, 156–157; 2006) could be adapted?

Most of us, most of the time, are solving simpler problems in biology. But we should always be alert to the great problems, such as understanding the processes leading to the origin of life. Given past experience, such great questions will be solved — not by chance but by the prepared mind. This book is a start to preparing that mind.

David Penny is at the Allan Wilson Center for Molecular Ecology and Evolution, Massey University, PO Box 11-222, Palmerston North, New Zealand.

Secret giants

Colossus: The Secrets of Bletchley Park's Codebreaking Computers

edited by B. Jack Copeland

Oxford University Press: 2006. 462 pp.

£18.99

Jon Agar

"I find it hard to remember in any detail, after keeping the secret for so long — I tried to blot out of my mind everything that happened at BP, for fear that I might talk in my sleep or under an anaesthetic," recalled Catherine Caghey, an operator of the Colossus, the wartime electronic cryptanalysis machine. The intense secrecy around Britain's codebreaking operations of the Second World War was only partially lifted in 1975. Brian Randell, investigating the gap in the life story of Alan Turing between the breathtaking response to Hilbert's decidability question in the 1930s and the construction of the first electronic stored-program computers in the late 1940s, was allowed to see a clutch of Colossus photographs. For another decade those grainy images, accompanied by terse texts, were the only historical trace of one of the most remarkable machines of the twentieth century.

Only in the 1990s did details emerge. Requests filed under the US Freedom of Information Act prised technical details, compiled by visiting liaison officers, from the American archives. Meanwhile, in Britain, veterans of Bletchley Park finally felt able to speak openly. No longer would they fear, like Caghey, that a trip to the dentist might result in a charge under the Official Secrets Act. Copeland's valuable collection brings together both technical commentary and personal narratives in an engaging book that will be essential reading for historians of twentieth-century technology and warfare. However, it is also clear that the secrecy of codebreaking did not merely conceal history, but shape it.

The Nazi state used many encryption machines. The Enigma, a cipher machine sold commercially in the 1930s, is now the most famous, a fame shared by Turing, the man most associated with cracking it. But

the higher echelons of Nazi Germany used a more sophisticated machine, the Lorenz SZ40, which was faster, entirely automatic (using Baudot–Murray teleprinter codes), and presumed invulnerable to attack. Eavesdropping stations across Britain recorded radio signals encrypted by the Lorenz SZ40. This stream of data, nicknamed Tunny, was channelled to Bletchley Park. Cracking Tunny was critical to allied success on D-Day: by reading decrypts it was known that Hitler expected only a feint at the Normandy beaches, while the delayed counterattack commands from the Führer to his generals could be read that day in London.

Bletchley Park struggled with Tunny. Two mathematical breakthroughs, by John Tiltman and William Tutte, gave the slenderest of wedges into decryption. But the essence of the problem was speed: only the freshest intelligence would save the most men and matériel. A creaking electromechanical machine, the Heath Robinson, helped in 1943. Fortunately, Tommy Flowers, a skilled Post Office engineer, was shown the problem. His proposal, a fully electronic cryptanalytical machine, met with scepticism even from senior codebreakers. But Flowers, with his team at the Post Office

research station at Dollis Hill in North London, built it anyway. Flowers triumphantly presented the Colossus to Bletchley Park in January 1944 (not, as had previously been recorded, in December 1943).

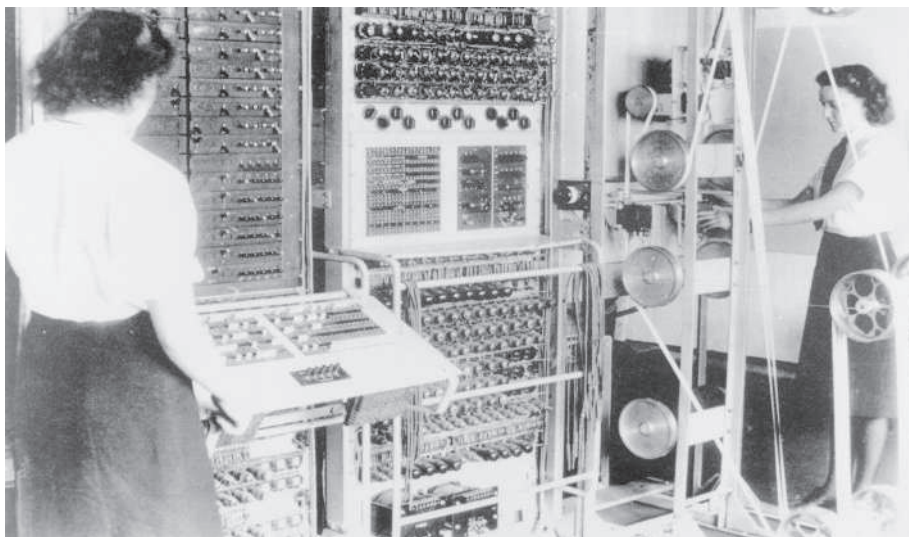
"I don't think they really understood what I was saying — I am sure they didn't," Flowers recalled, "because when the first machine was constructed and working, they were taken aback. They just couldn't believe it! ... I don't think they understood very clearly what I was proposing until they actually had the machine."

On 5 February 1944, Copeland writes, the "computer attacked its first message". Flowers wrote of the epochal event in his diary: "Colossus did its first job. Car broke down on way home".

But later he grew bitter: "When, after the war ended, I was told that the secret of Colossus was to be kept indefinitely, I was naturally disappointed. I was in no doubt, once it was a proven success, that Colossus was a historic breakthrough, and that publication would have made my name in scientific and engineering circles — a conviction confirmed by the reception accorded to ENIAC ... I had to endure all the acclaim given to that enterprise without being able to disclose that I had anticipated it."

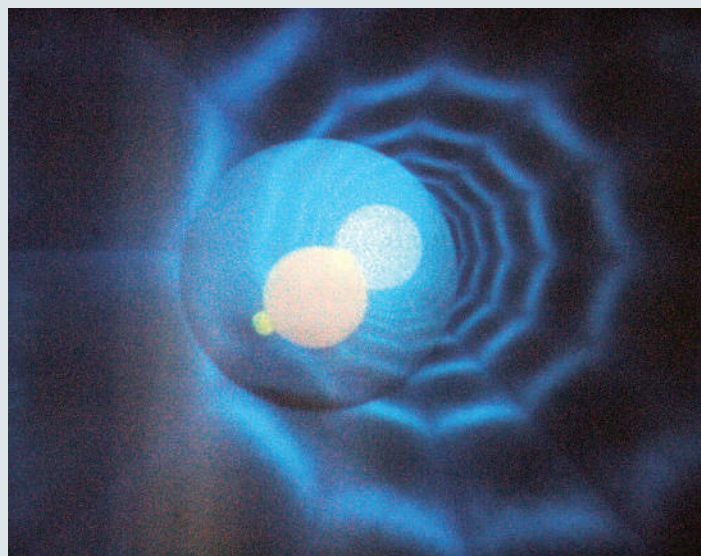
The team that worked on ENIAC, a high-speed electronic calculator in operation by 1945, conceived of stored programs. Neither ENIAC nor the Colossus was a computer, and it warps history to label them thus. But the key concept of the computer was publicized within months of the war's end. The British and US scientists who built the first electronic stored-program computers benefited from this openness. The Colossi, in contrast, were broken up, apart from two which ended up at the UK government's GCHQ listening centre at Cheltenham, and, for all intents and purposes, never appeared in the public world. Only one of these countries has sustained a thriving computer industry.

Jon Agar is in the Department of History and Philosophy of Science, University of Cambridge, Cambridge CB2 3RH, UK.



Cracking invention: the code-breaking Colossus helped Allied commanders intercept Hitler's messages.

BLETCHLEY PARK TRUST/SCIENCE & SOCIETY PICTURE LIBRARY



EXHIBITION

Discovering the invisible

What happens inside an accelerator when very high-energy particles collide? Visitors to the ALEXploratorium, the science centre within the Bibliotheca Alexandrina in Egypt, can now take a look for themselves (see picture).

An interactive exhibition, 'The Microscopes of Physics', introduces students and the broader public to both the smallest particles at the heart of atoms and the boundaries of the Universe. Successive rooms introduce visitors to the world

of atoms and their constituent particles, as revealed by particle accelerators and detectors, and on to the cosmos and its origins in the Big Bang.

The exhibition has been designed and brought about by the Italian National Institute for Nuclear Physics as part of the Mediterranean Association for Science Communication, in which science communicators from southern Europe and North Africa join forces to spread scientific culture. The exhibition runs until 31 October 2006.

Small science, big challenge

Nanotechnology: New Promises, New Dangers

by Toby Shelley

Zed Books: 2006. 208 pp. £45

Julia A. Moore

Toby Shelley's book *Nanotechnology* provides a short, accessible primer on the world of nanotechnology — the revolutionary realm of seeing, measuring, controlling and making things on the scale of atoms and molecules. It raises key questions about how this disruptive technology will affect human health, the environment, civil liberties, weaponry, and people in developing countries. Most important, it asks who decides which nanotechnology research projects will go forward, and who will profit from or be disadvantaged by the applications that result. It is big business: Lux Research reports that sales of products incorporating nanotechnology were worth US\$32 billion in 2005, and the US National Science Foundation predicts that nanotech-related goods and services could be a \$1-trillion market by 2015.

This book is long on high-minded questions. The author, a journalist at the *Financial Times*, supports some worthy and necessary steps. For example, he says that when governments evaluate the risks and benefits of nanotechnology, they should consider the health and environment impacts of nanotech products over their full life cycle, including development, manufacture, transport and disposal. He also says they should develop appropriate and internationally applied regulation to mitigate the risks without stifling innovation, and weigh carefully any military uses and encroachments on personal freedom.

On the downside, there is not much here that's new. All the analysis and prescriptions have been given before, sometimes with more

passion and insight. Much of the territory was covered by Bill Joy, the co-founder of Sun Microsystems, in his article 'Why the future doesn't need us' in *Wired* magazine in April 2000. And *Nanotechnology: A Gentle Introduction to the Next Big Idea* by Mark and Daniel Ratner (Prentice Hall, 2002) is a superior read. Shelley's book also contains some annoying minor errors and sweeping statements, particularly about other new technologies such as agricultural biotechnology, that many scientists will dispute.

In my view, nanotechnology will transcend today's Wall Street hype. It will eventually affect how people work, what they eat, how they communicate, and how long they live. It will change their medical care, energy sources, water and environment. As a result, its considerable benefits and possible risks need to be seriously and rigorously debated.

The progress made by science and technology continues to accelerate. Yet society seems increasingly incapable of answering questions, such as those raised in the 1970s by Alvin Toffler, about how we should respond to it. In 1995, the US Congress unwisely dismantled the Office of Technology Assessment (OTA), one of the world's few organizations to provide policy-makers with authoritative analyses of the complex scientific and technical challenges of the late twentieth century and beyond.

In the week when I finished reading Shelley's book, I attended the Enrico Fermi Award ceremony at the US National Academy of Sciences. The winner of this \$375,000 prize was Arthur Rosenfeld, who developed technological solutions and applied innovative policies aimed at reducing energy consumption.

A news feature published last year in *Nature* (438, 410–412; 2005) described what is increasingly known in energy circles as the

'Rosenfeld effect'. This refers to the ingenious policies and programmes devised by Rosenfeld that have helped California — the world's sixth-largest economy — keep per capita electricity consumption steady since the early 1970s, even though the state's economy was outpacing that of the United States as a whole. Today, California's per capita electricity use is the country's lowest, just over half that for the nation overall. Originally appointed by a Democrat, Rosenfeld was reappointed as a member of the California Energy Commission by Arnold Schwarzenegger, a Republican, when he became the state's governor.

Nanotechnologists need to find a way of cloning people such as Rosenfeld who are willing to roll up the sleeves of their lab coats and make good on science and engineering's promise to apply knowledge for the betterment of all mankind. The twenty-first century needs leaders like Benjamin Franklin who combine political and social vision with knowledge of science and engineering, and who understand how to move governments, public opinion, researchers and corporations to achieve true progress.

If nanotechnology is to be developed successfully, fresh kinds of policy insights, leadership and partnerships will be needed, alongside bold measures to create national and international assessment institutions that build on the OTA model. New global oversight mechanisms are required that can handle the complex convergence of nanotechnology, biotechnology, and information and cognitive science, and that are up to the task of 'taming the genie', as Shelley puts it in the final chapter.

Those looking for a roadmap to navigate this brave new world will find partial directions in Shelley's book. Hopefully, future books will help complete the journey. ■

Julia A. Moore is at the Woodrow Wilson International Center for Scholars, 1300 Pennsylvania Avenue, Washington DC 20004-3027, USA.

QUANTUM PHYSICS

A spin solo

Guido Burkard

Quantum computers could solve problems insurmountable to conventional computers. The missing ingredient for quantum computing with electron spins is now available — the rotation of a single spin.

How can one rotate the spin of a single electron? How can such a tiny rotation, once achieved, be detected? And where, indeed, does one obtain a single electron in the first place? The remarkable small semiconductor structures called quantum dots, whose tally of electrons can be tuned, one by one, down to just one¹, provide an answer to the third question. On page 766 of this issue², Koppens *et al.* supply a response to the first two. They exploit a phenomenon known as electron spin resonance to rotate a single spin in one of two coupled quantum dots, and detect the effect by measuring variations in electric current flowing through the double dot.

Electron spin resonance (ESR) results from the interplay of two magnetic fields. First, a static field is used to create an energy difference between an electron's spin-up and spin-down states, the phenomenon known as Zeeman splitting. Second, an oscillating field is applied at right angles to the first. When the frequency of this second field matches the frequency of the Zeeman energy gap, the electron's spin will rotate.

In a single quantum dot, if the Zeeman splitting were too small, thermal fluctuations in the contact leads would wash out the difference between the spin-up and spin-down energy states, preventing any reliable measurement of spin rotation. Thus, for ESR to occur and be clearly detected in a single dot, the Zeeman gap must typically have an energy equivalent to a high, microwave frequency.

ESR was first detected by the Soviet physicist Evgeny Zavoisky^{3,4} in the mid-1940s, and shortly thereafter in the United States⁵ and in Britain⁶. At that time, high-frequency sources — which came from the newly developed radar technology — were scarce, and microwave excitation remains a challenging business today. This is one reason why the analogous process of nuclear magnetic resonance (NMR) is so much more familiar: NMR spectroscopy typically requires a much lower, radio frequency, because the intrinsic magnetic moment of a nucleus (and therefore its Zeeman splitting) is about 2,000 times smaller than that of an electron.

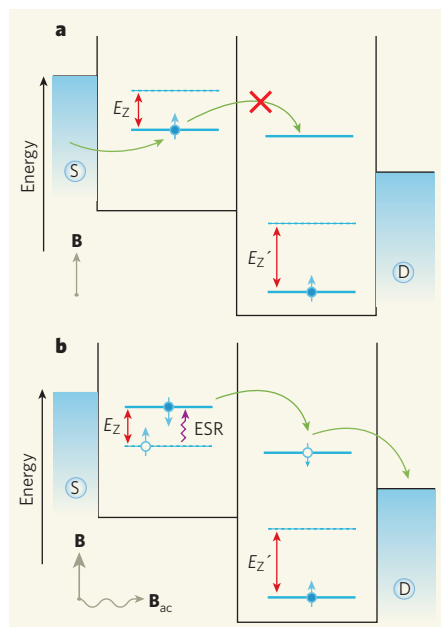


Figure 1 | Caught in the current. Koppens and colleagues² detect the rotation of a single electron spin induced by electron spin resonance (ESR) through measurements of electric current. **a**, A static applied magnetic field, B , causes the spin-up and spin-down energy levels of an electron to split (Zeeman splitting E_z and E_z'). The right-hand quantum dot (the deeper of the two wells, which has a larger splitting E_z') always contains at least one electron. An electron from a source (S) reservoir (a contact lead) hops on to the left dot. In about half of the cases, this electron has the same spin (up) as the electron on the right. Thus, the left-hand electron may hop no further, as the Pauli exclusion principle forbids two electrons with the same spin state from pairing up (a 'spin blockade'). **b**, A second, oscillating magnetic field B_{ac} initiates ESR, causing the left spin to flip. The right spin is unaffected, as it cannot muster the energy to jump the larger Zeeman gap E_z' in its well. Thus, right and left spins become antiparallel, and the blockade is lifted: the left electron hops via the right dot to the drain (D) reservoir, and a current is measured.

Koppens *et al.* circumvent the ESR frequency problem by measuring not the spin rotation in a single quantum dot directly, but how that rotation allows an electron to hop on to a second dot. This process is not affected by thermal fluctuations in the leads, so a smaller Zeeman splitting, which requires only radio-frequency excitation for ESR to occur, can be used.

Traditionally, ESR probes macroscopic samples by rotating roughly 10^{21} spins simultaneously. Rotating only one spin is a remarkable accomplishment by itself, but it is potentially also of great practical significance, because electron spins have been proposed as a way to realize 'quantum bits' — qubits — fundamental to a quantum computer⁷. A large-scale quantum computer might be able to master tasks, such as factorizing large numbers or simulating other quantum systems, that ordinary, classical computers cannot.

But quantum computers are delicate beasts, and very sensitive to disturbance from a noisy environment. Electron spins have the advantage as qubit candidates that they couple very weakly to their environment, by virtue of the

small electron magnetic moment. However, isolation from the environment is not enough for a good qubit (otherwise, we might well be discussing neutrino qubits). Crucially, spin qubits can be read out through converting spin information to charge information⁷, which can be detected at the single-electron level in quantum dots^{8,9}.

Additionally, spin qubits can be controlled by dint of a phenomenon known as exchange coupling, in which one spin can be used to affect the state of a neighbouring spin. This is a consequence of two fundamental physical laws: the Pauli exclusion principle, which prevents two electrons from assuming the same quantum state; and the repulsive Coulomb force between electrons' negative charges. Moreover, 'blank' spin qubits can be initialized reliably by illuminating the quantum dot with light from a laser¹⁰, a technique that also relies on the charge being coupled to the spin, in this case through the relativistic effect of spin-orbit coupling.

Five widely accepted 'hardware requirements' have been advanced for a practical quantum computer¹¹. These are: a scalable physical

system with well-characterized qubits; the ability to determine the initial values of those qubits; long-lived, coherent qubit states; the ability to measure single qubits; and a universal set of quantum operations that can be combined to generate an arbitrary algorithm, for example for factorization or quantum simulation. Of these requirements, all but the last have been realized for spin qubits in quantum dots.

All possible single-qubit operations together with the 'square-root-of-swap' on two qubits form a universal set appropriate for quantum computing⁷. The square-root-of-swap is a quantum operation that is rather like swapping the values of two qubits, but stopping midway — something that is impossible to do classically. Surprisingly, the two-qubit square-root-of-swap operation has already been demonstrated¹² in two exchange-coupled single-electron dots, and it is single-qubit operations that have languished as the last item on the 'to-do' list.

Using ESR for single-qubit operations was proposed early on¹³, but success has so far proved elusive, because detecting the phenomenon is far from straightforward. Koppens and colleagues² rely on the ingenious use of yet another sort of spin-to-charge conversion to detect the spin rotation indirectly. As originally suggested¹⁴, the idea was that the spin on a quantum dot can act like a valve for the electric current, which would repeatedly open and close when the spin is rotated. In two dots in series, this becomes a 'spin blockade'^{12,15}, in which the Pauli exclusion principle prevents current passing because otherwise electrons with parallel spins would be in the same place — and so in the same quantum state (Fig. 1).

Any process that can rotate one of the spins so that it is antiparallel can unblock this jam. Unblocking that is caused by the coupling of the electron spin in a quantum dot to the nuclear spins in its host semiconductor at low magnetic fields has already been observed¹². Koppens *et al.* set up the conditions for ESR to occur in their quantum-dot system and observe just such an unblocking behaviour. They observe both a continuous current when the ESR field is turned on constantly, and so-called Rabi oscillations of the current (and thereby the spin) for a pulsed ESR field with varying pulse length.

The observation of a current is not enough on its own to prove that ESR has caused the rotation of a single spin. The oscillating magnetic field that sets up ESR also, unavoidably, brings with it a small electric field. Through a process known as photon-assisted tunnelling, this field can assist electrons in breaking the spin blockade by, for instance, first hopping back into the source contact, against the applied voltage. This leads to a false alarm: electric current in the absence of spin rotation. Although it turns out that a double dot is less susceptible to such processes than a single dot, Koppens and colleagues further suppress it by increasing the energy barrier for

back-hopping by applying a large voltage between the two contacts. Direct suppression of the oscillating field was achieved by applying a voltage to one of the contacts that oscillated with the same frequency, but opposite phase, compared with the undesired electric field.

The ability to rotate a single electron spin in a quantum dot, combined with single-spin read-out and controllable exchange coupling between spins, throws up many questions and challenges. Prominent among these is the coupling of the electron spin to the surrounding nuclear spins, which reduces the accuracy of spin rotations. Sequences of several subsequent pulses instead of one, as widely used in NMR, might improve the situation, but in the long run one also needs to gain sufficient control over the nuclear spins. That turns out to be an interesting problem in itself^{16–19}.

When such difficulties have been surmounted, the stage is set for first tests of quantum algorithms with a small, but increasing number of spin qubits. The tool-box for spin-based quantum computing is complete — now it's time to use it. ■

Guido Burkard is in the Department of Physics and Astronomy, University of Basel, Klingelbergstrasse 82, CH-4056 Basel, Switzerland.
e-mail: guido.burkard@unibas.ch

1. Ciorga, M. *et al.* *Phys. Rev. B* **61**, R16315–R16318 (2000).
2. Koppens, F. H. L. *et al.* *Nature* **442**, 766–771 (2006).
3. Zavoisky, E. J. *Phys. USSR* **9**, 211–216 (1945).
4. Zavoisky, E. J. *Phys. USSR* **10**, 197–198 (1946).
5. Cummerow, R. L. & Halliday, D. *Phys. Rev.* **70**, 433 (1946).
6. Bagguley, D. M. S. & Griffiths, J. H. E. *Nature* **160**, 532–535 (1947).
7. Loss, D. & DiVincenzo, D. P. *Phys. Rev. A* **57**, 120–126 (1998).
8. Elzerman, J. M. *et al.* *Nature* **430**, 431–435 (2004).
9. Hanson, R. *et al.* *Phys. Rev. Lett.* **94**, 196802 (2005).
10. Atatüre, M. *et al.* *Science* **312**, 551–553 (2006).
11. DiVincenzo, D. P. *NATO ASI Ser. E* **345**, 657–677 (1997).
12. Petta, J. R. *et al.* *Science* **309**, 2180–2184 (2005).
13. Burkard, G., Loss, D. & DiVincenzo, D. P. *Phys. Rev. B* **59**, 2070–2078 (1999).
14. Engel, H.-A. & Loss, D. *Phys. Rev. Lett.* **86**, 4648–4651 (2001).
15. Ono, K., Austing, D. G., Tokura, Y. & Tarucha, S. *Science* **297**, 1313–1317 (2002).
16. Giedke, G. *et al.* preprint available at www.arxiv.org/quant-ph/0508144 (2005).
17. Stepanenko, D. *et al.* *Phys. Rev. Lett.* **96**, 136401 (2006).
18. Klausner, D., Coish, W. A. & Loss, D. *Phys. Rev. B* **73**, 205302 (2006).
19. Lai, C. W. *et al.* *Phys. Rev. Lett.* **96**, 167403 (2006).

GENOMICS

Predictable packaging

Timothy J. Richmond

Nuclear factors must access specific sites within genomic DNA to function, yet the DNA is bundled up into many nucleosomes. Is the DNA sequence sufficiently informative to predict where each nucleosome will be?

The genomes of eukaryotic organisms — plants, fungi and animals — range in length from 10 million to 100 billion base pairs. All of this DNA is packaged into a structure known as chromatin so that it can fit into the cell nucleus, which is only a few micrometres in diameter. The nucleosome, the fundamental repeating unit of chromatin, consists of 160–240 base pairs of DNA (depending on the species and tissue type) wrapped around a core of proteins called histones. So there are roughly 50,000 to 500 million nucleosomes per cell. On page 772 of this issue, Segal *et al.*¹ attempt the mammoth task of predicting the positions of all the nucleosomes in a genome, based solely on DNA sequence.

Why bother? The answer is that DNA not only encodes the instructions for life in its sequence of base pairs, but also provides signposts for regulating its own readout. These signpost elements enable regulatory factors to take up precise locations in the genome that are essential for their function. For example, a gene-activating factor might bind to the segment of DNA immediately next to its target gene.

Nucleosomes play several roles in such factor recognition. Most obviously, they must

be positioned so that authentic functional sites associated with initiation of gene activity are exposed. Conversely, they must mask spurious recognition sites that occur coincidentally in large genomes. These are the nucleosome's static roles. In addition, their positioning might block individual sites required for stepwise formation of functional multiprotein complexes, sliding away on demand with the aid of energy-dependent, chromatin-remodelling factors. To completely package an expanse of DNA, shutting its genes down entirely, nucleosomes adopt an even denser, higher-order organization in which the DNA is further concealed (Fig. 1). The resulting chromatin fibre structure requires a certain periodicity of nucleosome positions along the DNA.

Consideration of these issues suggests that nucleosomes are unlikely to exist randomly along the DNA. It might be that factor binding occurs on newly replicated DNA before chromatin assembly, thereby influencing nucleosome location, but there is scant evidence for this theory. Nucleosome mapping experiments demonstrate clearly that nucleosomes *in vivo* and *in vitro* take up preferred positions with respect to DNA sequence. How, then,

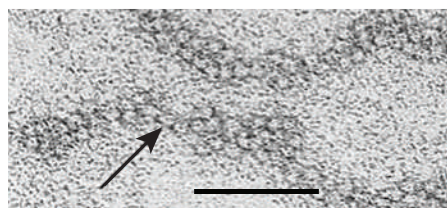


Figure 1 | Nucleosome organization. Chromatin fibres showing nucleosomes (small circles). Regions containing a single nucleosome (arrow) may represent transitions between different packing arrangements. Segal *et al.*¹ have shown that the locations of the nucleosomes may be determined directly by the DNA sequence. Scale bar, 0.1 μm . (Reproduced from ref. 10.)

does DNA sequence contribute to nucleosome positioning? And can we predict these positions for entire genomes simply by analysing the DNA sequence?

Two decades ago, Satchwell *et al.*² discovered that there are significant periodicities of pairs of adjacent DNA bases (dinucleotides, combinations of the four possible bases A, T, C and G) in DNA from nucleosomes isolated from chickens. Numerous analytical and predictive studies of nucleosome position followed, with the main conclusion being that the sequence-dependent bending flexibility of DNA is a significant factor^{3,4}.

Segal *et al.*¹ raise the bar for the prediction of nucleosome position. They have developed a free-energy function (analogous to a partition function in statistical thermodynamics) that evaluates the probability of a nucleosome occurring at any base pair in a genome and taking into account occlusion by neighbouring nucleosomes — an important consideration, given the close proximity of nucleosomes and their packing in chromatin fibres. As input, the authors mapped 199 yeast nucleosomes and determined the frequency of dinucleotides in the 147 base pairs that are tightly wrapped up in each nucleosome core⁵. They found, as did Satchwell *et al.*², that AA/TT/TA dinucleotides are favoured where both DNA-strand phosphodiester backbones (the side-rails of the DNA ladder) face inward, towards the histone-protein core. GC is favoured where both backbones face outward.

The probabilities determined by Segal *et al.* could predict nucleosome position with roughly 50% accuracy, substantially better than the 35% expected by chance. But how do they know when they are right? To answer this, the authors used two types of data. First, they gathered nucleosome positions mapped *in vivo* from published studies. These studies typically focus on gene regulatory regions, but one study addressed the entire yeast genome⁶. The data are generally limited in positional accuracy to about ± 30 base pairs, so each of the authors' predictions was scored 'correct' if it was within ± 35 base pairs of a mapped position.

Second, Segal *et al.* evaluated nucleosome affinities for yeast DNA both by selecting nucleosomes *in vitro* based on stability, and by

identifying high-occupancy nucleosomes in the yeast genome using a technique known as real-time PCR. Limiting the prediction to these nucleosomes yielded better correspondence: there was a sixfold improvement over random expectation using the set of nucleosomes selected *in vitro*.

Where in genomes are the most stable nucleosomes? Strikingly, it seems that in genes containing the TATA sequence (which signals the starting point for gene expression) they are immediately downstream of the TATA sequence, but not engulfing it. Overall, the greatest number of high-occupancy nucleosomes occur between genes (and so are probably regulatory in function), whereas nucleosomes of the greatest stability are found in centromeres — the most tightly packed region of the chromosome. Conversely, nucleosomes tend not to inhabit the functional DNA-binding sites for gene regulatory factors, as these sites have lower than average occupancy. It seems, therefore, that a large contrast between high and low affinity occurs in regulatory regions, suggesting that there is a high level of organization. One further remarkable feature is that correlated positioning of neighbouring nucleosomes extends over at least six nucleosomes, yielding a repeat of 177 base pairs in yeast. Such long-range organization would promote chromatin fibre formation⁷.

What limits the prediction? First, the accuracy of the mapping is inadequate. A crucial improvement would be the availability of DNA array chips possessing a resolution down to 3 base pairs and covering the entire yeast genome as well as substantial parts of larger eukaryote genomes. Second, dinucleotide frequencies are probably insufficient as input. Although they partially represent the bending flexibility of DNA, the structure of the nucleosome core particle reveals that tetranucleotides and possibly longer DNA sequences are involved⁵. Third, the actions of chromatin remodelling factors and subsequent competition by DNA-binding factors will result in perturbations in the positions of some nucleosomes. With further improvements in prediction, these differences will lead to a better description of the mechanism of gene activation.

Lastly, there is the enigmatic histone H1 question. The H1 protein binds to the linker DNA that connects nucleosome cores, and promotes compaction of the chromatin fibre. When nucleosomes are assembled on arrays *in vitro*, their spacing depends critically on the concentration of salt or polycationic H1 (ref. 8). Recently, Fan *et al.*⁹ reduced the amount of H1 associated with chromatin *in vivo* by deleting half of the H1-encoding genes in three lines of mouse embryonic stem cells. They showed that the spacing between nucleosomes drops by 30 base pairs in all cases. Despite this, the cells survive and the chromatin remains compact. Thus, the DNA sequence may not be the most significant determinant of nucleosome position, after all.



50 YEARS AGO

Physics and Microphysics by Prof. Louis de Broglie — The late Albert Einstein contributes an all too brief foreword to this book, but he aptly sums up the remarkable series of essays which it contains by referring to it as a "unique book". Louis de Broglie is primarily a theoretical physicist, and it was through his careful and critical analysis, in the early 1920's, of the necessity to reconcile the granular and undulatory aspects of light that he was led to suggest the extension of the idea of the duality of waves and corpuscles to all elements of matter... de Broglie is a modest man, but it is necessary for him to look back in retrospect on the origins of wave mechanics; thus, Chapter 8, which is devoted to personal memories on the beginning of wave mechanics, is excellent reading and shows up clearly the deep insight to physical problems that de Broglie possesses... He concludes the last chapter, "The Great Adventure", with the words: "In the work of science man has been able to show the force of his intelligence; if he wishes to survive his own successes, it is necessary that he should now show the wisdom of his will". This is a book that every physicist should read and ponder over.

From *Nature* 18 August 1956.

100 YEARS AGO

The Origin and Development of the Moral Ideas by Dr. Edward Westermarck — In one engaging paragraph of this work, its author describes how, whilst living in the North of Morocco — where he spent four years studying folklore — he was described as a person with "propitious ankles," because the village where he stayed was frequently visited by favoured and distinguished guests. Propitiousness is not with us the most familiar term in such a context, but the ankles of Dr. Westermarck's intellectual endeavour are certainly sturdy.

From *Nature* 16 August 1906.

50 & 100 YEARS AGO

This apparent contradiction can be circumvented by assuming that the weak, 10-base-pair sequence periodicity found in the DNA of the nucleosome core permits 10-base-pair displacements without incurring large energy penalties. The results of Segal *et al.*¹ and Fan *et al.*⁹ suggest that nucleosome arrays could have fixed phase origins of periodicity that may or may not depend solely on the properties of DNA. Comparative mapping of at least a fraction of the genomes for the normal

and H1-deleted cell lines should reveal these putative, unperturbed sites of highly stable nucleosomes.

Timothy J. Richmond is at the ETH Zürich, Institut für Molekularbiologie und Biophysik, Schafmattstrasse 20, 8093 Zürich, Switzerland.

e-mail: richmond@mol.biol.ethz.ch

1. Segal, E. *et al.* *Nature* **442**, 772–778 (2006).
2. Satchwell, S. C., Drew, H. R. & Travers, A. A.

J. Mol. Biol. **191**, 659–675 (1986).

3. Anselmi, C., Bocchinfuso, G., De Santis, P., Savino, M. & Scipioni, A. *Biophys. J.* **79**, 601–613 (2000).
4. Widom, J. *Q. Rev. Biophys.* **34**, 269–324 (2001).
5. Richmond, T. J. & Davey, C. A. *Nature* **423**, 145–150 (2003).
6. Yuan, G. C. *et al.* *Science* **309**, 626–630 (2005).
7. Schalch, T., Duda, S., Sargent, D. F. & Richmond, T. J. *Nature* **436**, 138–141 (2005).
8. Blank, T. A. & Becker, P. B. *J. Mol. Biol.* **252**, 305–313 (1995).
9. Fan, Y. *et al.* *Cell* **123**, 1199–1212 (2005).
10. Ratner, J. B. & Hamkalo, B. A. *J. Cell Biol.* **81**, 453–457 (1979).

SOLID-STATE PHYSICS

Resistance is futile

Adam C. Durst

With the right combination of microwave radiation and magnetic field, two-dimensional electron systems conduct electricity with zero resistance. But is this zero really zero, or is it negative resistance in disguise?

Electrons confined to the interface between two semiconductors form what is known as a two-dimensional electron gas — essentially, a two-dimensional metal. Thanks to advances in semiconductor technology over the past few decades, two-dimensional electron gases of remarkably high purity can now be fabricated. These materials provide a playground for the exploration of fundamental physics, and have made possible the discovery of many important phenomena. Writing in *Physical Review Letters*, Zudov *et al.*¹ describe a clever experiment designed to shed light, quite literally, on one such phenomenon: ‘zero-resistance’ states induced by the application of a magnetic field and microwave radiation.

The resistance of a material measures the degree to which it impedes the flow of an electrical current. In a metal, electrons flow easily, so metals have relatively low resistance. Indeed, if the crystal lattice of atoms that define the metal were perfectly periodic, the electrons would flow with zero resistance. In practice, materials have imperfections from which electrons can scatter, disrupting the flow of current. Thus, real metals have some resistance.

Under the right conditions, however, certain materials actually can exhibit zero resistance. This occurs not because there is no scattering, but because all the electrons work together to form a collective state of matter that is immune to the effects of scattering on current flow. Famous examples of such phenomena include superconductivity and the magnetic-field-induced quantum Hall effects.

Another system with zero resistance was discovered^{2–4} when microwave radiation was shone on a two-dimensional electron gas in a magnetic field far too weak for quantum Hall effects to be observed. The microwaves cause the resistance of the sample to oscillate as a

function of the applied magnetic field. The period and amplitude of these oscillations are controlled by the frequency and intensity of the applied radiation, respectively. For high enough intensity, the oscillation amplitude is large enough for the minima to reach all the way down to zero resistance, where they go no further — they saturate at zero, in what have been called zero-resistance states (Fig. 1a).

A logical question is whether this phenomenon is yet another example of zero resistance indicating a collective state of matter. So far, the consensus seems to be no. Soon after the experimental discovery, theorists^{5–8} provided a relatively simple explanation of the microwave-induced resistance oscillations. Essentially, the absorption of microwaves serves to push electrons either with or against their normal direction of flow (which way depends on the relationship between the frequency of the microwave radiation and the frequency at which electrons orbit in the applied magnetic field). In one case this increases the electrical resistance and in the other case it decreases it.

For radiation of sufficient intensity, calculations showed that the minima of the resistance oscillations would pass right through zero resistance, and a regime of negative resistance would result. Negative resistance is a perfectly reasonable concept for a system injected with energy in the form of microwave radiation: a laser works in a similar way, achieving negative attenuation (gain) when pumped with energy. However, a two-dimensional electron gas with uniform current flow and negative resistance was soon shown to be unstable⁹: such a system prefers to form an inhomogeneous current pattern characterized by zero resistance. Thus, at present, the theoretical picture is that the zero-resistance states indicate not a new collective state, but rather an example of pattern formation in a system out of equilibrium.

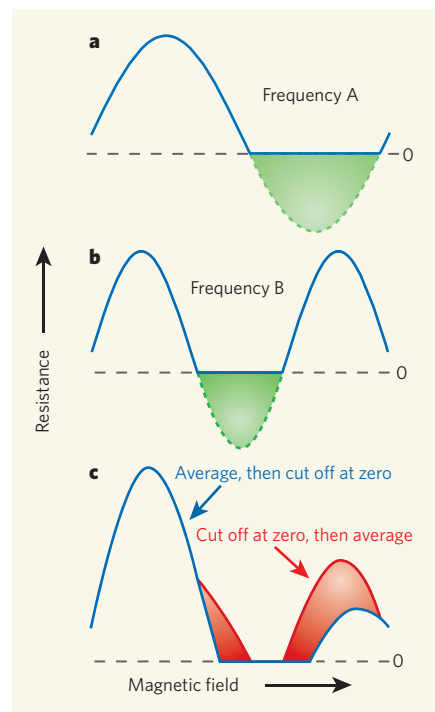


Figure 1 | Unmasking negative resistance.

a, b, Schematic depictions of microwave-induced resistance oscillations and zero-resistance states for applied radiation of two distinct frequencies. Minima of the oscillations are cut off at zero resistance. **c**, If microwaves of both frequencies are applied at once, it is unclear how their effects should be combined. If the zeros are fundamental, each curve should be cut off at zero first, and then averaged (red curve). If the zeros are masking negative resistance, the full oscillations should be averaged first (including the dashed negative parts), and the result cut off at zero (blue curve). The measurements of Zudov *et al.*¹ support the second approach.

Zudov *et al.*¹ set out to test this assertion by determining whether the zero-resistance state fundamentally has zero resistance, or whether it denotes a regime where the resistance would be negative if negative resistance were not unstable. Their approach amounts to the following. Average the ‘zero’ with a positive resistance R . If the result is $R/2$, the ‘zero’ was really zero. If the result is less than $R/2$, the ‘zero’ was really a negative resistance in disguise.

The authors address their task by subjecting their sample to radiation of two different

frequencies at once. If applied individually, radiation of each frequency would result in microwave-induced resistance oscillations and zero-resistance states, but with maxima and minima at different values of the magnetic field (Fig. 1a, b). For high-intensity radiation, it is reasonable to imagine that the combined effect of both frequencies will be an average of the two curves. If the zeros are fundamental, we should cut off each curve at zero, and then average (red curve in Fig. 1c). But if the zeros are masking negative resistance, we should average the two curves first, taking into account any negative regions, and then cut off the result at zero (blue curve in Fig. 1c). In magnetic-field regimes where one curve is positive and the other is zero, the second of these two procedures will yield a smaller result than the first. Zudov and colleagues' measurements¹ agree with what one would obtain from the second procedure.

These results support (but by no means prove) the idea that this particular zero-resistance state emerges because the system is desperate to avoid a state of negative resistance, instead choosing to form a rather unusual,

inhomogeneous current configuration. Imaging these current patterns — which have yet to be observed directly — and developing a theoretical understanding of their structure¹⁰ are exciting directions for research now and in the future.

Adam C. Durst is in the Department of Physics and Astronomy, Stony Brook University, Stony Brook, New York 11794-3800, USA. e-mail: adam.durst@stonybrook.edu

1. Zudov, M. A., Du, R. R., Pfeiffer, L. N. & West, K. W. *Phys. Rev. Lett.* **96**, 236804 (2006).
2. Zudov, M. A., Du, R. R., Simmons, J. A. & Reno, J. L. *Phys. Rev. B* **64**, 201311 (2001).
3. Mani, R. G. *et al.* *Nature* **420**, 646–650 (2002).
4. Zudov, M. A., Du, R. R., Pfeiffer, L. N. & West, K. W. *Phys. Rev. Lett.* **90**, 046807 (2003).
5. Durst, A. C., Sachdev, S., Read, N. & Girvin, S. M. *Phys. Rev. Lett.* **91**, 086803 (2003).
6. Ryzhii, V. I. *Sov. Phys. Solid State* **11**, 2078–2080 (1970).
7. Ryzhii, V. I., Suris, R. A. & Shchamkhalova, B. S. *Sov. Phys. Semicond.* **20**, 1299–1302 (1986).
8. Dmitriev, I. A. *et al.* *Phys. Rev. B* **71**, 115316 (2005).
9. Andreev, A. V., Aleiner, I. L. & Millis, A. J. *Phys. Rev. Lett.* **91**, 056803 (2003).
10. Auerbach, A., Finkler, I., Halperin, B. I. & Yacoby, A. *Phys. Rev. Lett.* **94**, 196801 (2005).

ASTRONOMY

Young spirals get older

Robert C. Kennicutt Jr

These days, galaxies come in very different shapes and sizes. Cutting-edge technologies allow a detailed peek at how things looked in the Universe's early days — 'the same, but different' is the tentative message.

On page 786 of this issue¹, Genzel *et al.* present remarkable observations of what appears to be a newly formed spiral galaxy, observed when the Universe was just a fifth of its current age. The result is doubly significant: first, it provides the most detailed glimpse so far of the formation of a galaxy similar to our own Milky Way; second, it demonstrates the power of a new generation of high-resolution instruments that use adaptive optics to study the information and evolution of far-off galaxies.

Astronomers have made exceptional progress in recent years in identifying distant young galaxies and tracing their evolution over cosmic time. The farther away a galaxy is, the longer its light takes to get to us. So by studying galaxies at ever larger distances, we can observe the Universe at an ever younger age. Over the past few years, thousands of distant galaxies that trace back billions of years, some to as little as a tenth of the Universe's current age, have been identified.

These observations have given rise to a self-consistent picture of the evolution of galaxies. The present-day 'Hubble sequence' division into elliptical and spiral galaxies (Box 1, overleaf) was already clear when the Universe was about a third to a half as old as it is now. When

we look back to even earlier times, however, we see that the nature of the galaxy population becomes dramatically different. Here, it is dominated by compact young objects, and by clumped fragments of stars and gas at an earlier stage in their evolution that will eventually transform into fully formed galaxies.

These observations are in rough accord with the current hierarchical theory of galaxy formation. But so far, large spiral galaxies with well-developed disks similar to the Milky Way are conspicuously absent in both observations and models of the early Universe. These large spirals are expected to form rather late, so one would not expect to find many of them at early times, and the diffuse light of a young disk is hard to detect at large distances. Spectroscopic instruments attached to the world's largest telescopes, which have apertures of 8–10 metres, are now being used to identify candidate objects and study their structure and internal dynamics^{2,3}. But even these instruments are handicapped by the relatively poor image quality available from the ground. Above Earth's atmosphere, the Hubble Space Telescope can produce much sharper images, but lacks the aperture — and currently a spectrograph — to measure such faint objects.

Genzel *et al.*¹ break new ground by using adaptive optics, in which the light of a bright star close to the object of interest is used to correct for image distortions caused by Earth's atmosphere. One of the European Southern Observatory's four 8-metre-aperture Very Large Telescopes, situated at the La Silla Paranal site in northern Chile, was used as a light-gathering tool, and the resulting images have a resolution rivaling that of Hubble Space Telescope pictures. The authors fed the corrected images into a specially built spectrometer to obtain a simultaneous picture of emissions at infrared wavelengths in 1,024 regions of 32×32 pixels.

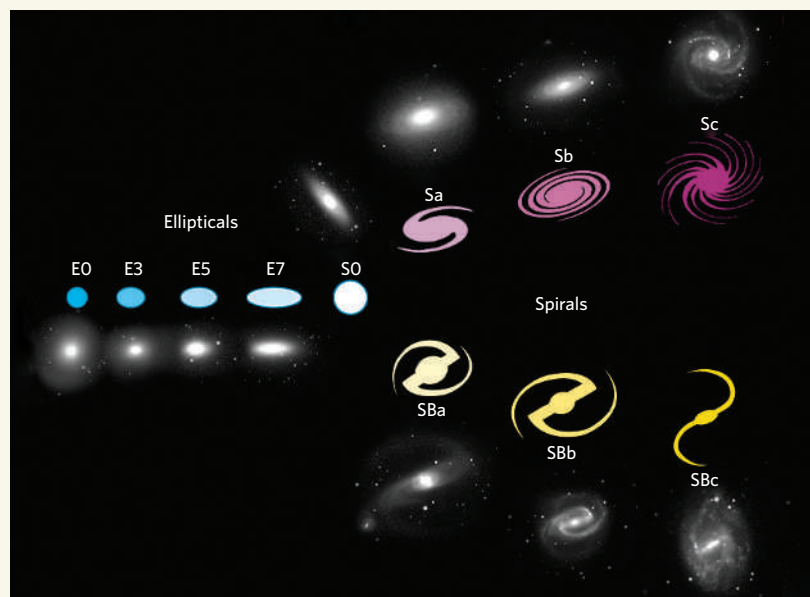
The object of this technical outlay was a luminous, star-forming galaxy known as BzK-15504. This galaxy, whose exact form had remained unresolved from previous observations, was chosen for its location next to a bright foreground star that could be used as a beacon for adaptive-optics corrections. Because this galaxy has been moving away from us as the Universe expands, its light is redshifted by a factor of 2.38; the infrared light mapped by the spectrograph is thus actually redshifted visible light — the so-called hydrogen Balmer- α emission at 656 nm — emitted by the galaxy 10 billion years ago. This radiation is produced by ionized gas clouds surrounding star-forming regions in the galaxy, and can be used to trace the distribution and motions of the gas.

The authors' observations of BzK-15504 reveal it to be a giant spiral galaxy, with a size and mass similar to that of the Milky Way, but observed just 3 billion years after the Big Bang. It shows many similarities to present-day spiral galaxies, with rotational properties that, again, are nearly identical to those of the Milky Way. These similarities are notable because they imply that at least some large disk galaxies were broadly in place even at these early cosmic epochs.

But other properties of BzK-15504 indicate a young, turbulent galaxy that might still be in the final stages of formation. By contemporary standards, the galaxy is creating stars at a stupendous rate — equivalent to about 140 Suns a year. That's about 50 times higher than the present rate of star formation in the Milky Way, and ten times higher than that in the most massive present-day spiral galaxies. Such high rates can only be sustained for a small fraction of a galaxy's lifetime, so we are probably witnessing a transient burst of star formation caused by an accretion or merger event, or perhaps an inaugural spurt initiated by the formation of the disk itself.

Irregular motions of BzK-15504's galactic disk exposed in the Balmer- α data¹ might provide supporting evidence for accretion or merger processes. But such peculiarities could also be induced by energy releases from exploding stars — supernovae — associated with the intense star formation. Even more tantalizing are hints that gas might be flowing

Box 1 | The Hubble sequence and the evolution of structure



Edwin Hubble's 'tuning fork' classification divides galaxies into two principal categories. First, there are the relatively featureless elliptical galaxies. These are subdivided into categories from E0 to E7 with increasing 'eccentricity' (deviation from a spherical form). Elliptical galaxies tend to consist of older stars, show very little active star formation, and are dominated by random motion.

Second, spiral galaxies — our own Milky Way is an example — consist of a central 'bulge', resembling an elliptical galaxy, surrounded by a flat, rotating disk of younger stars in a characteristic arm formation. Spiral galaxies are broken down into barred and normal spirals, and classified according to how tightly the arms of the disk are wound around the bulge.

Observations of galaxies at early times in the Universe's evolution present an opportunity to

see how the Hubble sequence looked back then. The widely accepted hierarchical theory of structure formation presents a 'bottom-up' picture, in which large galaxies grow by the accretion of smaller fragments and by mergers. This gravitational clumping is guided by the unseen hand of so-called cold dark matter^{5,6}.

Simulations based on the hierarchical model reproduce impressively the large-scale structure of the Universe and compact protogalaxies seen at early cosmic times. These protogalaxies then evolve into elliptical galaxies and the central bulges of spiral galaxies. But the models have difficulty producing more than a few large, spiral galaxy disks — especially at early times. Genzel and colleagues' observation¹ of a spiral galaxy that seems well on its way to being fully formed at an early epoch is thus a considerable challenge to these established models. **R.C.K.**

into the centre of the galaxy. That could fuel a burst of star formation there, or perhaps even trigger an energetic eruption from a central black hole. Drawing clear conclusions from a single object is difficult, but further observations will reveal how widespread such processes are in young spiral galaxies. Both these and other results² from the same programme are challenging theorists to account for the existence of such massive and well-formed galaxies at such early cosmic epochs.

What impresses most about these results is that in-depth studies of the structure and the internal dynamics of young galaxies are now demonstrably within our reach. The next technological step will be the mating of adaptive-optics systems with laser guide stars⁴ — a laser beam shot up into the sky where no appropriate star is available for calibration of atmospheric effects. That should allow these kinds of observations to be extended to virtually any galaxy in the sky. Such facilities are under construction or being commissioned on several of the world's largest telescopes, and observations of young galaxies are central to the case for the planned next generation of observatories with 20–40-metre-aperture telescopes. Just one adolescent spiral galaxy has shown what the scientific potential of those facilities might be. ■

Robert C. Kennicutt Jr is at the Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK.
e-mail: robk@ast.cam.ac.uk

1. Genzel, R. *et al. Nature* **442**, 786–789 (2006).
2. Förster-Schreiber, N. M. *et al. Astrophys. J.* **645**, 1062–1075 (2006).
3. Erb, D. K. *et al. Astrophys. J.* **646**, 107–132 (2006).
4. Wizinowich, P. L. *et al. Publ. Astron. Soc. Pacif.* **118**, 297–309 (2006).
5. Springel, V. *et al. Nature* **435**, 629–636 (2005).
6. Springel, V., Frenk, C. S. & White, S. D. M. *Nature* **440**, 1137–1144 (2006).

CANCER BIOLOGY

A game of subversion

Emmanuelle Passegué

Just as stem cells are crucial for tissue development and regeneration, cancer stem cells underlie tumour formation and maintenance. But do cancer stem cells invariably arise from normal stem cells?

It is now clear that certain tumours can be sustained by a rare population of cancer stem cells, which share one of the defining properties of normal stem cells — the ability to renew themselves. Self-renewal is what allows stem cells to persist during the lifetime of the organism and to provide new cells for tissue genesis, maintenance and regeneration following stress or injury. These properties are exactly what cancer stem cells exhibit in initiating and maintaining malignant growth

— and, unfortunately, in regenerating a tumour when the cancer stem cells escape treatment.

A key question therefore is whether cancer stem cells are always derived from normal stem cells that run amok, producing cancer cells instead of normal cells (Fig. 1a, b). This idea is certainly plausible because many cancer-stem-cell populations express cell-surface proteins that are also found on normal stem cells. But stem cells might not be the only

source of cancer stem cells. On page 818 of this issue, Krivtsov and colleagues¹ report the generation of cancer stem cells by a leukaemia-associated protein that has found a way to trick non-stem cells into acquiring stem-cell-like behaviour and supporting tumour formation (Fig. 1c)*.

Human leukaemias often harbour recurrent chromosome translocations, where segments of the chromosomes have been moved around. For instance, the 'mixed lineage leukaemia' gene (*MLL*) is repeatedly found fused to various partners, including the *AF9* gene used by Krivtsov *et al.*¹. *MLL* is the vertebrate homologue of the fruitfly *trithorax* gene that regulates many developmental programmes by maintaining expression of the *Hox* family of genes², which are master regulators of development. *MLL* fusion genes are associated with

*This article and the paper concerned¹ were published online on 16 July 2006.

the development of acute-type myeloid leukaemias³. In mice, they can initiate leukaemia when expressed either in the blood-forming stem cells or, more relevant to this study, in myeloid progenitors only^{4,5}. Myeloid progenitors are cells that are produced by the blood-forming stem cells and that have begun to specialize, so they do not self-renew and are normally destined solely to give rise to mature white blood cells or myeloid cells⁶.

Cellular differentiation is usually considered to be a one-way process of specialization as the cells develop the functions of their ultimate fate and lose their immature characteristics such as self-renewal. So how does the fusion gene 'transform' myeloid progenitors into cancer-forming cells, given that these progenitor cells have seemingly gone past the stage of self-renewal? Krivtsov *et al.*¹ answer this question by showing that *MLL-AP9* induces the aberrant expression of a specific set of 'stem cell genes' in myeloid progenitor cells. This makes the cells self-renew and turns them into cancer stem cells capable of initiating, maintaining and propagating the leukaemia.

Self-renewal is not yet fully understood at the molecular level, but it can be tracked using microarrays to profile the expression levels of hundreds of genes⁷. Hence, stem-cell identity can be summed up as a gene-expression signature that differs from the signature of non-stem-cell populations. Some of these differences are likely to underpin self-renewal and other stem-cell-specific properties. What Krivtsov *et al.*¹ show is that transformed myeloid progenitors aberrantly express a small number of stem-cell genes (363 genes), while still displaying the overall gene-expression profile of myeloid progenitor cells. So the transformed progenitors do not become stem cells but rather acquire stem-cell-like behaviour. The authors show that the acquired stem-cell signature is lost when the cancer stem cells are forced to give up their self-renewal properties. This result strongly suggests that the stem-cell genes identified do indeed mediate the aberrant self-renewal of the myeloid progenitor cells.

But what are these genes and how can they make non-stem cells self-renew? This is the real million-dollar question, and Krivtsov *et al.*¹ have begun to address it. First, they provide evidence that the stem-cell self-renewal programme is hierarchical in nature, with a specific set of genes activated as an early response to *MLL-AP9* expression. This early programme includes known regulators of

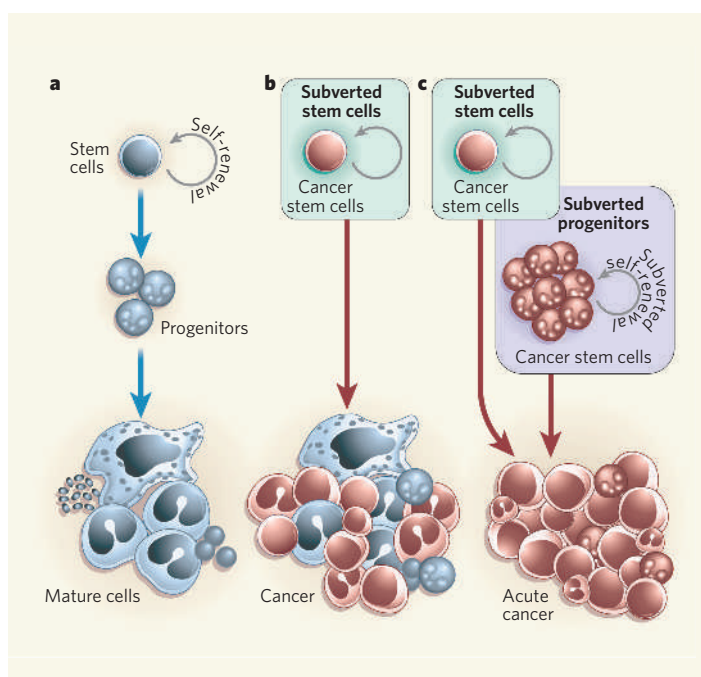


Figure 1 | Sources of cancer stem cells. **a**, Normal stem cells are the only cells that self-renew for the lifetime of the organism, and they produce the progenitor and mature cell populations required for tissue function and maintenance (here, the blood system). **b**, Cancer genes can subvert the stem-cell population to turn it into cancer stem cells, which then produce cancer cells instead of normal cells. **c**, Krivtsov *et al.*¹ find that some cancer genes involved in acute myeloid leukaemia can also turn progenitor cells into cancer stem cells by inducing expression of a specific set of stem-cell genes. This expression subverts the progenitor cells to self-renew and to support tumour formation.

stem-cell self-renewal activity, such as members of the *Hox* gene family⁸. Second, the authors examine one of these early-response genes, *Mef2c*, and show that it modulates stem-cell behaviour *in vitro* and *in vivo*. The overall architecture of the self-renewal programme still needs to be defined, as do the specific genes and/or pathways that are crucial for its establishment both in normal- and cancer-stem-cell populations. However, Krivtsov and colleagues' results bring hope that interfering with such deregulated self-renewal genes may prevent or reverse the formation of cancer stem cells.

One concern in the cancer field is whether acquisition of stem-cell features in non-stem cells occurs in human diseases or whether it is an artefact of using the mouse as a disease model. Krivtsov *et al.*¹ weigh in to this long-standing debate by showing that part of the mouse stem-cell self-renewal signature is expressed in cells isolated from human patients suffering from leukaemia associated with *MLL* fusion genes. Although the presence of contaminating normal stem cells within these leukaemia specimens cannot be excluded, these findings support the idea that subversion of stem-cell properties in non-stem-cell populations can happen in the course of a human disease.

Is this really so unexpected? A fundamental property of cancer cells is the ability to subvert the normal mechanisms that restrain unfettered growth. The loss of self-renewal

potential during normal differentiation might be viewed as such a safeguard, so it follows that uncoupling the property of self-renewal from other stem-cell characteristics could have potent cancerous effects. *MLL* is a particularly strong candidate for this sort of sabotage as it normally regulates stem-cell-fate decisions by maintaining the expression of *Hox* genes, which in turn support self-renewal^{2,8}. It is now necessary to find out whether *MLL* fusion genes can transform human progenitor cells, either directly or as a consequence of transforming stem cells, and whether this occurs by a true activation of self-renewal genes rather than by a lack of gene repression. It is also necessary to determine whether this type of subversion occurs with other fusion genes in leukaemia and with other types of cancer.

Krivtsov and colleagues' findings are a step towards defining what could be a universal self-renewal signature for human leukaemia. Such a signature would clearly have

a bearing on cancer treatment, not only enabling targeting of the cancer stem cells that emerge from non-stem-cell populations, but also identifying the self-renewal mechanisms that can go awry in normal stem cells to make them cancerous⁹. Cancer genes play a subtle game of subversion that we could ultimately use against them. By appropriating mechanisms that are essential to the functions they want to gain (here self-renewal), cancer genes teach us how these processes work, and give us the opportunity to identify molecular targets that could be used for cancer treatment. ■

Emmanuelle Passegué is in the Developmental and Stem Cell Biology Program and the Department of Medicine, University of California, San Francisco, 513 Parnassus Avenue, San Francisco, California 94314, USA.
e-mail: passegué@stemcell.ucsf.edu

1. Krivtsov, A. V. *et al.* *Nature* **442**, 818–822 (2006).
2. Ernst, P., Mamon, M., Davidson, A. J., Zon, L. I. & Korsmeyer, S. J. *Curr. Biol.* **14**, 2063–2069 (2004).
3. Daser, A. & Rabbitts, T. *Genes Dev.* **18**, 965–974 (2004).
4. Cozzio, A. *et al.* *Genes Dev.* **17**, 3029–3035 (2003).
5. So, E. *et al.* *Cancer Cell* **3**, 161–171 (2003).
6. Akashi, K., Traver, D., Miyamoto, T. & Weissman, I. L. *Nature* **404**, 193–197 (2000).
7. Forsberg, E. C. *et al.* *PLoS Genet.* **1**, e28 (2005).
8. Abramovich, C. & Humphries, R. K. *Curr. Opin. Hematol.* **12**, 210–216 (2005).
9. Passegué, E., Jamieson, C. H. M., Ailles, L. E. & Weissman, I. L. *Proc. Natl Acad. Sci. USA* **100**, 11842–11849 (2003).

OBITUARY

George W. Wetherill (1925–2006)

Geochemist, planetary scientist and astrobiologist.

Theories of the origin of the Solar System are necessarily uncertain, as the experiment that led to its formation cannot be repeated. But in the past few decades, our idea of how Earth and the other planets of the Solar System came to be has advanced from highly speculative musing to a true scientific theory with a firm base in hard facts. George W. Wetherill, who died on 19 July, was a leader in this effort to explain Earth's origins.

Wetherill was born in Philadelphia, Pennsylvania, on 12 August 1925. After service in the Second World War, during which he taught the principles of radar at the Naval Research Laboratory in Washington DC, he studied at the University of Chicago, receiving a PhD in nuclear physics there in 1953. He then moved to the Department of Terrestrial Magnetism at the Carnegie Institution of Washington, where he began work on the first major challenge of his career: how to date geological samples precisely.

At that time, the department's physicists, led by Merle Tuve, were exploring diverse subjects in Earth science, including the use of radioactive decay to date the point of crystallization of terrestrial rocks. In this environment, Wetherill was able to estimate for the first time the half-life of the extremely long-lived radioactive rubidium isotope ^{87}Rb by determining the rate at which ^{87}Rb must have decayed since its rocks formed — an age obtained independently by measurements of uranium decay. That meant that rocks could be dated precisely through the relative abundance of rubidium and its decay product, strontium. Together with potassium–argon dating, this technique allowed precise ages to be determined from a far wider range of rocks, such as granites, that did not contain the rarer uranium ores.

Wetherill was also able to bring clarity to data on the lead isotopes that, as decay products of uranium, had puzzled geochemists for years with their discordant ages. Wetherill developed the concordia diagram as a reliable means of obtaining precise crystallization ages, as well as evidence for later metamorphic disturbances, from uranium–lead isotopic ratios. The diagram found immediate, lasting acceptance among geochemists, and Wetherill's injection of modern physics into geochemistry opened up the new field of radiometric geochronology.

In 1960, partly to avoid the domineering Tuve, Wetherill accepted the offer of a professorship at the University of California, Los Angeles, where he undertook his second great challenge: understanding the orbital

dynamics of asteroids and comets, and the likelihood that they might hit Earth. The switch in focus from laboratory geochemistry to theoretical calculations of orbits was entirely consistent with Wetherill's lifelong desire to work only where he could make a significant contribution to solving great outstanding problems.

Here too, he met with success, determining the precise chance that a body orbiting in the asteroid belt between Mars and Jupiter would end up on a collision course with Earth. These calculations provided a strong indication that meteorites found on Earth originated in the asteroid belt. Thus, laboratory measurements of the elemental and isotopic composition of meteorites provided information on the make-up of bodies elsewhere in the Solar System that were otherwise inaccessible to detailed analysis. Wetherill also showed that the highest-velocity fragments of bodies hitting Mars might be launched on orbits that would intersect that of Earth, a prediction spectacularly verified by the discovery of dozens of martian meteorites on Earth.

Wetherill returned to the Department of Terrestrial Magnetism at Carnegie as its director in 1975. He was a fine director, but treated it as a part-time post. Anyone approaching him in his office would inevitably find him scanning a computer printout or a screen detailing the latest results of his own research, and he would only reluctantly divert his attention to departmental business. If it was a really serious matter, he would spin his ever-present pencil lengthwise between his fingers as he considered what to do. His first response to any request was invariably 'no'. But if one troubled to come back a second time, he would view the matter as important enough to perhaps reconsider the request. On the whole, the staff liked having a director who was more interested in doing his own work than in interfering with theirs.

At Carnegie, Wetherill embarked on his third major task — clarifying the origin of Earth itself. Here, he built on the pioneering work of the Soviet scientist Victor Safronov, whose 1969 monograph on planet formation had become widely available in a translated version only in 1972. Wetherill adapted his calculations of the orbital dynamics of asteroids to compute the evolution of small, randomly colliding rocky bodies orbiting the early Sun. Using so-called Monte Carlo methods and increasingly powerful digital computers, he showed that a swarm of hundreds of lunar-mass bodies could form



one or two Earth-mass planets (that is, Venus and Earth), and a few smaller rocky planets (Mercury and Mars), orbiting at roughly the observed distances from the Sun.

This process would probably also involve one or two Mars-mass bodies hitting the growing Earth. If such a giant impact occurred off-centre, a spray of hot mantle would be placed in Earth orbit. That debris would later coalesce into the Moon. This neat solution eliminated problems that had bedevilled previous theories of lunar origin, and, like the concordia diagram, the giant-impact theory rapidly found widespread acceptance.

Wetherill continued his research after retiring in 1991, calculating how terrestrial planets might form around other stars, and showing that Earth-like planets could form with masses several times that of Earth. If that is so, the extrasolar planets of less than ten Earth masses so far discovered might be just the high-mass outliers of a host of Earth-like worlds, with significant implications for the existence of life elsewhere in the Universe. Many details remain to be explored theoretically and tested by observations of other planetary systems, but this could be Wetherill's greatest legacy: a scientifically robust, universal theory of the formation of habitable, Earth-like planets.

Alan P. Boss

Alan P. Boss is in the Department of Terrestrial Magnetism, Carnegie Institution of Washington, 5421 Broad Branch Road NW, Washington DC 20015-1305, USA.
e-mail: boss@dtm.ciw.edu

CARNEGIE INST.

BRIEF COMMUNICATIONS

Silent spread of H5N1 in vaccinated poultry

A chink in the protection of a caged flock can dramatically increase the chances of a flu outbreak.

International debate on the merits of vaccinating poultry against the H5N1 influenza A virus^{1–3} has raised concerns about the possibility of an increased risk of between-flock transmission before outbreaks are detected⁴. Here we show that this 'silent spread' can occur because of incomplete protection at the flock level, even if a vaccine is effective in individual birds. The use of unvaccinated sentinels can mitigate, although not completely eliminate, the problem.

We use an individual-based mathematical model, parameterized from experimental and observational data⁵ (for details, see supplementary information), that tracks within-flock spread of a highly pathogenic avian influenza (HPAI) virus such as H5N1 to explore the impact of prophylactic vaccination on silent spread between flocks. We calculated the probability of outbreak occurrence and detection, as well as the contribution of flock infectiousness during, and at the end of, a flock's production cycle to transmission to other flocks (expressed as the case reproduction ratio, R_{between}). We determined the quantitative effects on these variables of different flock structure, within-flock transmission potential of HPAI, detection thresholds, vaccine effectiveness and fraction of the flock successfully vaccinated.

Outbreaks were modelled in caged flocks seeded with a small amount of infective faeces contaminating a single cage. The fraction of birds successfully vaccinated was considered⁶, assuming a fully effective vaccine. We find that 90% of birds need to be protected to reduce the probability of an outbreak by 50% (Fig. 1a, solid line), but this can result in undetected outbreaks (Fig. 1a, blue dashed line). The infectiousness of an infected flock to other flocks (defined as the infectiousness of faeces integrated hourly over an outbreak) during the production cycle peaks at 80% protection (Fig. 1b, solid line). This is because, as the fraction protected rises, fewer birds become infected but outbreaks become harder to detect. Despite a reduced probability of outbreaks, vaccination can increase between-flock transmission at high flock-protection levels (Fig. 1b, dashed

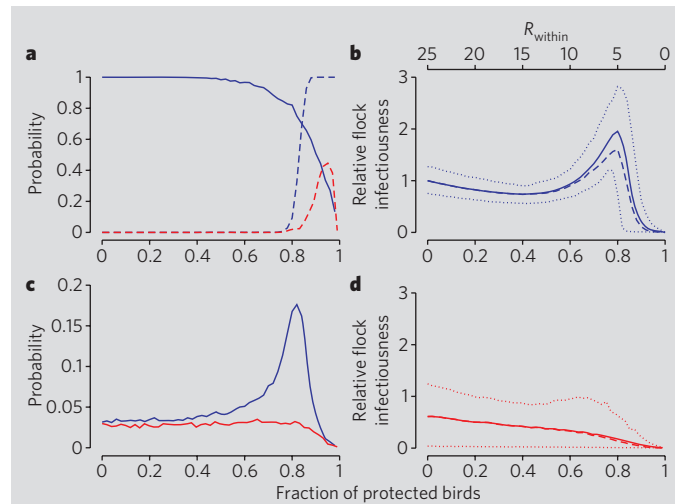


Figure 1 | Effects of vaccination protection levels in a flock. **a**, Probability of an outbreak (solid line) and probability that an outbreak will ever be detected (dashed lines) in caged flocks of 10,000 birds. Line colours: blue, no sentinels; red, 100 sentinels. **b**, Mean (solid line) and 95 percentiles (dotted lines) of the infectiousness of infected vaccinated flocks during the production cycle, relative to infected unvaccinated flocks (no sentinels). Dashed line: infected flock infectiousness weighted by probability of outbreak occurrence. For reference, the corresponding values of within-flock transmission potential, R_{within} , are shown. **c**, Probability that an outbreak occurs and is undetected at the end of the production cycle, assuming a cycle period of 365 days. Line colours: blue, no sentinels; red, 100 sentinels. **d**, Flock infectiousness as in **b**, but with 100 sentinels.

line). These results are qualitatively robust to uncertainties in parameter values (for details, see supplementary information).

The risk of between-flock transmission is greatest at the end of a flock's production cycle, when biological security can be compromised as birds are moved and housing units cleaned. High levels of flock protection can dramatically increase the probability that HPAI is undetected at this time because of the increased outbreak duration (Fig. 1c, blue line), thus contributing to between-flock transmission.

The negative effects of vaccination can be mitigated by monitoring unvaccinated sentinel birds placed into flocks⁴ (although logistical problems arise⁷). Sentinels placed randomly among cages increase the probability of detection (although undetected outbreaks still occur; Fig. 1a, dashed red line), thereby reducing flock infectiousness (Fig. 1d) and the probability of undetected HPAI at the end of a production cycle (Fig. 1c, red line).

Between-flock transmission is related to flock infectiousness both during (Fig. 1b, d) and at the end of (Fig. 1c) a production cycle. Values of R_{between} as high as 3 to 10 have been

reported⁸. To prevent an epidemic, R_{between} must be below 1 (ref. 9): to achieve this through vaccination, even with sentinels, would require very high levels of flock protection (Fig. 1c, d, red lines).

The main obstacle to achieving such protection is in ensuring that an adequate fraction of birds is properly vaccinated⁷ (typically less than 90% of birds are protected in practice²). Vaccination can be highly effective in individual birds¹⁰ and any minor deficiencies at that level are relatively unimportant (see supplementary information). A successful vaccination programme therefore requires not only a highly effective vaccine but also a highly effective vaccine-delivery system, combined with effective biological security and the rapid detection and removal of infected flocks.

Nicholas J. Savill*,
Suzanne G. St Rose*,
Matthew J. Keeling†,
Mark E. J. Woolhouse*

*Centre for Infectious Diseases,
University of Edinburgh,
Ashworth Laboratories,

The King's Buildings, Edinburgh EH9 3JT, UK
e-mail: nick.savill@ed.ac.uk

†Mathematics Institute and Department of
Biological Sciences, University of Warwick,
Coventry CV4 7AL, UK

1. Marangon, S., Bortolotti, I., Capau, M., Bettio, M. & Dalla Pozza, M. *Avian Dis.* **47**, 1006–1009 (2003).
2. Ellis, T. M. et al. *Avian Pathol.* **33**, 405–412 (2004).
3. Butler, D. *Nature* **434**, 810 (2005).
4. European Commission Vaccination of Poultry against Highly Pathogenic Avian Influenza H5N1 (DIVA Strategy) (2006); at http://ec.europa.eu/food/animal/diseases/controlmeasures/avian/discussion_paper.pdf
5. Elbers, A. R. et al. *Avian Dis.* **48**, 691–705 (2004).
6. Woolhouse, M. E., Haydon, D. T. & Bundy, D. A. *Vet. J.* **153**, 41–47 (1997).
7. Suarez, D. L. *Biologicals* **33**, 221–226 (2005).
8. Stegeman, A. et al. *J. Infect. Dis.* **190**, 2088–2095 (2004).
9. Woolhouse, M. et al. *Nature* **411**, 258–259 (2001).
10. Garcia, A. et al. *Avian Dis.* **42**, 248–256 (1998).

Supplementary information accompanies this communication on Nature's website.

Received 31 March; accepted 12 July 2006.

Competing financial interests: declared none
doi:10.1038/442757a

BRIEF COMMUNICATIONS ARISING online
www.nature.com/bca see Nature contents.

CELL BIOLOGY

Nondisjunction, aneuploidy and tetraploidy

Arising from: Q. Shi & R. W. King *Nature* 437, 1038–1042 (2005)

One simple, widely accepted mechanism for generating an aberrant chromosome number, or aneuploidy, is through nondisjunction — a chromosome distribution error that occurs during mitosis when both copies of a duplicated chromosome are deposited into one daughter cell and none into the other. Shi and King¹ challenge this view, concluding that nondisjunction does not yield aneuploid cells directly, but instead gives rise to tetraploid cells that may subsequently become aneuploid through further division. Here we show that the direct result of chromosome nondisjunction is gain or loss of a single chromosome, which results in near-diploid aneuploidy, not tetraploidy. We suggest that chromatin trapped in the cytokinetic cleavage furrow is the more likely reason for furrow regression and tetraploidization.

Shi and King use fluorescent *in situ* hybridization to establish a correlation between spontaneous missegregation of chromosomes and binucleation in human cell lines¹. Notably, they find the same correlation in cells with an intact mitotic checkpoint as in cells in which the checkpoint has been inactivated as a result of depletion of the essential checkpoint protein Mad2. We tested whether nondisjunction *per se* affects completion of mammalian cytokinesis by using recombinase-mediated gene excision to deplete the mitosis-specific motor protein CENP-E from primary fibroblasts. Although mitotic checkpoint signalling is weakened at individual CENP-E-depleted kinetochores, mitosis proceeds as normal in most of these cells, with all duplicated chromatid pairs achieving alignment and then segregation in anaphase. In 25% of divisions, however, one or two duplicated chromosomes fail to make bipolar attachments. These cells sustain a mitotic checkpoint-mediated delay for about an hour, but then anaphase ensues and nondisjunction occurs by retention of both copies of individual, duplicated chromosomes at or near one of the spindle poles² (Fig. 1a).

If nondisjunction were to give rise to binucleation and if binucleate cells were to survive as well as mononucleate cells for one cell cycle, as demonstrated by Shi and King¹, then the steady-state proportion of binucleates in the population of CENP-E-depleted cells should triple. Instead, examination of over 500 cells of each genotype revealed no difference in the proportion of binucleation in CENP-E-depleted cells and in wild-type cells (Fig. 1b), which have a nondisjunction rate below 0.5%.

Examination of other cell types from mice supports our conclusion that nondisjunction

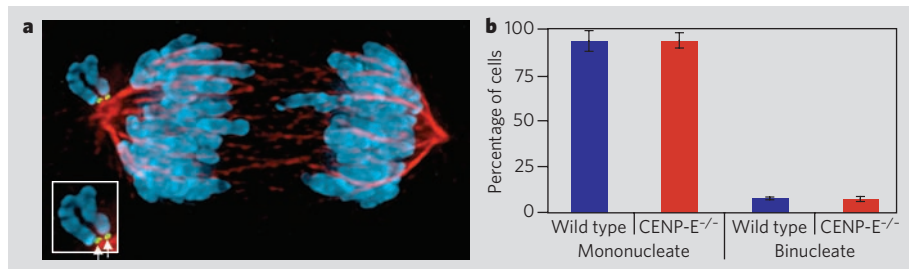


Figure 1 | Nondisjunction causes near-diploid aneuploidy, not cytokinesis failure, in primary cells.

a, Anaphase chromosome distribution in a primary mouse fibroblast depleted of CENP-E (after CENP-E gene disruption mediated by Cre recombinase²). Inset, micrograph enlargement of one pair of duplicated, sister chromosomes undergoing nondisjunction. Arrows, centromeres. DNA, blue; tubulin, red; centromeres of unattached chromosomes (BUB1), green. **b**, Percentages of binucleate cells in wild-type and CENP-E-null populations; $n > 500$ cells from each of two independent experiments.

Methods. Chromosome spreads from lymphocytes from peripheral blood were prepared according to the Jackson Laboratory protocol (www.jax.org/cyto/blood_preps.html). The predicted percentage of binucleation in CENP-E-null cells was calculated as follows: if nondisjunction drives binucleation, the total number of binucleate progeny produced per division for CENP-E-null cells will be $0.25n$, where n is the number of mononucleated cells. As Shi and King have shown that binucleates survive (and either exhibit an extended interphase arrest or divide to produce daughters that are, in some cases, also binucleate)¹, and as long-term survival occurs in binucleate cells produced by inhibiting cytokinesis^{8–10}, most binucleates should survive for at least the length of one cell cycle. The total number of cells at the next division then is $0.25n + 2 \times 0.75n$. The steady-state proportion of binucleates generated by nondisjunction would thus be $0.25n / (0.25n + 2 \times 0.75n)$, or 14%, plus the 7% rate derived from other mechanisms in the parental wild-type cells, yielding an overall 21% binucleates, if nondisjunction yields binucleation.

causes near-diploid aneuploidy. Chromosome counting revealed aneuploidy in 28% of mitotic lymphocytes from mice with reduced levels of CENP-E (compared with 10% in normal mice). This situation is similar in mice with reduced amounts of the mitotic checkpoint protein Bub3, in which 9% of splenocytes are aneuploid compared with fewer than 1% in normal mice³. In both cases, almost all instances of aneuploidy reflect the loss or gain of only one or two chromosomes, exactly as would be expected to result directly from simple nondisjunction.

Shi and King also propose that nondisjunction-induced failure of cytokinesis, coupled with tetraploidization and a subsequent aberrant mitosis, underlies the aneuploidy frequently found in human cancer. However, nondisjunction occurs at high frequency in mosaic variegated aneuploidy, a condition characterized by childhood cancers and thought to be caused by mutations in the mitotic checkpoint kinase BubR1 (refs 4–6); this produces trisomy or monosomy of individual chromosomes at high frequency in multiple tissues^{5,6}. Thus, although single chromosome gain or loss in cancer is rare and cytokinesis failure has probably occurred in some tumours that are highly aneuploid, current evidence indicates that nondisjunction produces near-diploid aneuploidy in primary

mammalian cells, mice and humans.

A possible explanation for the correlation observed by Shi and King¹ is that one or more mitotic defects occurred that, in addition to causing nondisjunction, resulted in trapping of DNA in the cytokinetic furrow. Indeed, the authors observe chromatin bridging or lagging (events likely to lead to DNA trapping in the cleavage furrow) in more than half of cells that subsequently became binucleate. Although they conclude that there was no chromosome bridging in the remainder, some mitotic errors producing DNA within the cleavage furrow, including the strands of DNA that result from stretching of single kinetochores that are attached to microtubules from both poles, were below their sensitivity of detection. In our view, then, the most likely explanation for furrow regression and the subsequent binucleate, tetraploid state is not nondisjunction, but rather a DNA-damage response caused by chromatin breakage by the incoming cytokinetic furrow — as first suggested almost thirty years ago by the evidence in ref. 7.

Beth A. A. Weaver, Alain D. Silk, Don W. Cleveland

Ludwig Institute for Cancer Research and Department of Cellular and Molecular Medicine, University of California, San Diego, La Jolla, California 92093-0670, USA
e-mail: dcleveland@ucsd.edu

1. Shi, Q. & King, R. W. *Nature* **437**, 1038–1042 (2005).
2. Weaver, B. A. A. *et al.* *J. Cell Biol.* **162**, 551–563 (2003).
3. Babu, J. R. *et al.* *J. Cell Biol.* **160**, 341–353 (2003).
4. Matsuura, S. *et al.* *Am. J. Med. Genet. A* **140**, 358–367 (2006).
5. Hanks, S. *et al.* *Nature Genet.* **36**, 1159–1161 (2004).
6. Limwongse, C., Schwartz, S., Bocian, M. & Robin, N. H. *Am. J. Med. Genet.* **82**, 20–24 (1999).
7. Mullins, J. M. & Bieseke, J. J. *J. Cell Biol.* **73**, 672–684 (1977).
8. Andreassen, P. R., Lohez, O. D., Lacroix, F. B. & Margolis, R. L. *Mol. Biol. Cell* **12**, 1315–1328 (2001).
9. Uetake, Y. & Sluder, G. *J. Cell Biol.* **165**, 609–615 (2004).
10. Wong, C. & Stearns, T. *BMC Cell Biol.* **6**, doi:10.1186/1471-2121-6-6 (2005).

doi:10.1038/nature05139

CELL BIOLOGY

Shi & King reply

Replying to: B. A. A. Weaver, A. D. Silk & D. W. Cleveland *Nature* **442**, doi:10.1038/nature05139 (2006).

Weaver *et al.* report¹ that chromosome non-disjunction induced by knockout of the mitosis-specific motor protein CENP-E does not increase the rate of binucleation in mouse fibroblasts and conclude that nondisjunction is therefore not coupled to furrow regression. However, they have not yet excluded alternative interpretations of their results, which makes it hard to compare them directly to our findings. Critical confounding issues include analysis of different cell types and the use of a specific gene deletion, for which corresponding naturally occurring mutations have not been described.

First, the differences could be explained by cell-type-specific differences in cytokinesis regulation. We evaluated three human epithelial cell lines², but the mouse fibroblasts examined by Weaver *et al.* may generate greater traction forces³ that could promote cytokinesis completion through cell tearing⁴, rather than by the targeted delivery and fusion of vesicles near the midbody⁵. Furrow regression occurs many hours after mitosis in HeLa cells² and, during the intervening period, cells remain connected by a thin cytoplasmic bridge. In fibroblasts, increased traction forces could break this bridge, uncoupling chromosome nondisjunction from furrow regression. Analysis of cytokinesis completion in CENP-E^{+/+} and CENP-E^{-/-} mouse fibroblasts by time-lapse imaging is necessary to exclude this possibility. Furthermore, it is essential that the CENP-E removal experiment be repeated in epithelial cells, which give rise to most human cancers.

Second, conclusions based solely on the

analysis of CENP-E^{-/-} cells is complicated by the fact that CENP-E itself (and its binding partner, BubR1) may be involved in coupling chromosome nondisjunction and cytokinesis completion. CENP-E is required for efficient congression of mono-orientated chromosomes to the metaphase plate⁶ and for spindle checkpoint signalling⁷. Following anaphase, CENP-E localizes to the midbody⁸ where it may negatively regulate cytokinesis completion⁹. Thus, given the many important functions of CENP-E in mitosis, CENP-E could act directly in coupling nondisjunction to cytokinesis completion, explaining why CENP-E removal would uncouple these two processes.

Third, nondisjunction in CENP-E^{-/-} cells is unlike spontaneous nondisjunction. Nondisjunction induced by loss of CENP-E results from cells that enter anaphase with chromosomes remaining at the poles¹⁰. In contrast, we identified 17 cells that underwent nondisjunction, yet only two cells showed defects in chromosome congression². Spontaneous nondisjunction therefore seems to be a different process from nondisjunction occurring in CENP-E^{-/-} cells, providing another potential explanation for differences in the subsequent regulation of cytokinesis. The causes of spontaneous nondisjunction are poorly understood, but our time-lapse analysis indicates that perturbation of CENP-E function is unlikely to be among them.

Weaver *et al.* suggest that furrow regression is due to the presence of DNA in the cleavage furrow¹, but this possibility is inconsistent with our data: of 395 cells that completed cytokinesis,

58 showed chromosome bridging or lagging, indicating that these defects are not sufficient to induce furrow regression². Furthermore, 47% of cells that became binucleated showed no bridging or lagging, indicating that these defects are not required for furrow regression. Although thin strands of trapped chromatin may be beyond our limit of detection, these cells also showed no evidence of micronuclei or disrupted interphase nuclear architecture, which might arise as a result of such defects. In contrast, cytokinesis failure has been reported only in cells showing gross distortions in nuclear architecture, with teardrop-shaped nuclei connected to one another by a large chromatin bridge¹¹. We observed defects of this magnitude in only 11% of cells that became binucleated. Thus, although bridging and lagging may be associated with binucleation in a fraction of cells, we conclude that other mechanisms must exist to link nondisjunction and furrow regression in the remaining cells.

In budding yeast, there is a pathway that delays cytokinesis completion to prevent damage to chromosomes near the division site¹². Although this mechanism has not been demonstrated in animal cells, our results indicate that such a pathway could possibly be activated more generally by spontaneous nondisjunction.

Qinghua Shi*, **Randall W. King†**

*Hefei National Laboratory for Physical Sciences at Microscale, School of Life Sciences, University of Science and Technology of China, Hefei, Anhui 230026, China

†Department of Cell Biology, Harvard Medical School, Boston, Massachusetts 02115, USA
e-mail: randy_king@hms.harvard.edu

1. Weaver, B. A. A., Silk, A. D. & Cleveland, D. W. *Nature* **442**, doi:10.1038/nature05139 (2006).
2. Shi, Q. & King, R. W. *Nature* **437**, 1038–1042 (2005).
3. Burton, K. & Taylor, D. L. *Nature* **385**, 450–454 (1997).
4. Piel, M., Nordberg, J., Euteneuer, U. & Bornens, M. *Science* **291**, 1550–1553 (2001).
5. Gromley, A. *et al.* *Cell* **123**, 75–87 (2005).
6. Kapoor, T. M. *et al.* *Science* **311**, 388–391 (2006).
7. Putkey, F. R. *et al.* *Dev. Cell* **3**, 351–365 (2002).
8. Yen, T. J. *et al.* *EMBO J.* **10**, 1245–1254 (1991).
9. Liu, D. *et al.* *Biochem. Biophys. Res. Commun.* **345**, 394–402 (2006).
10. Weaver, B. A. A. *et al.* *J. Cell Biol.* **162**, 551–563 (2003).
11. Mullins, J. M. & Bieseke, J. J. *J. Cell Biol.* **73**, 672–684 (1977).
12. Norden, C. *et al.* *Cell* **125**, 85–98 (2006).

doi:10.1038/nature05140

REVIEWS

Multiferroic and magnetoelectric materials

W. Eerenstein¹, N. D. Mathur¹ & J. F. Scott²

A ferroelectric crystal exhibits a stable and switchable electrical polarization that is manifested in the form of cooperative atomic displacements. A ferromagnetic crystal exhibits a stable and switchable magnetization that arises through the quantum mechanical phenomenon of exchange. There are very few ‘multiferroic’ materials that exhibit both of these properties, but the ‘magnetoelectric’ coupling of magnetic and electrical properties is a more general and widespread phenomenon. Although work in this area can be traced back to pioneering research in the 1950s and 1960s, there has been a recent resurgence of interest driven by long-term technological aspirations.

Since its discovery less than one century ago, the phenomenon of ferroelectricity¹, like superconductivity, has been considered in relation to the ancient phenomenon of magnetism. Just as recent work has shown that magnetic order can create superconductivity², it has also been shown that magnetic order can create (weak) ferroelectricity³ and vice versa^{4,5}. Single-phase materials in which ferromagnetism and ferroelectricity arise independently also exist, but are rare⁶. As this new century unfolds, the study of materials possessing coupled magnetic and electrical order parameters has been revitalized. In this Review we set recent developments in the context of the pioneering works of the 1950s–1960s.

The field of research that we are describing has a tortuous taxonomy and typically involves terms such as ‘multiferroic’ and ‘magnetoelectric’, whose overlap is incomplete (Fig. 1). By the original definition, a single-phase multiferroic⁷ material is one that possesses two—or all three—of the so-called ‘ferroic’ properties: ferroelectricity, ferromagnetism and ferroelasticity (Fig. 2 and Table 1; see Box 1 for a glossary of terms). However, the current trend is to exclude the requirement for ferroelasticity in practice, but to include the possibility of ferrotoroidic order (Box 1) in principle. Moreover, the classification of a multiferroic has been broadened to include antiferroic order (Box 1). Magnetoelectric coupling (Box 1), on the other hand, may exist whatever the nature of magnetic and electrical order parameters, and can for example occur in para-

magnetic ferroelectrics⁸ (Fig. 1). Magnetoelectric coupling may arise directly between the two order parameters, or indirectly via strain. We also consider here strain-mediated indirect magnetoelectric coupling in materials where the magnetic and electrical order parameters arise in separate but intimately connected phases (Fig. 3).

A confluence of three factors explains the current high level of interest in magnetoelectrics and multiferroics. First, in 2000, Hill (now Spaldin) discussed the conditions required for ferroelectricity and ferromagnetism to be compatible in oxides, and declared them to be rarely met⁶. Her paper in effect issued a grand materials development challenge that was taken up because empirically there are indeed few multiferroic materials, whatever the microscopic reasons. Second, the experimental machinery for the synthesis and study of various contenders was already in place when this happened. Third, the relentless drive towards ever better technology is aided by the study of novel materials. Aspirations here include transducers and magnetic field sensors, but tend to centre on the information storage industry.

It was initially suggested that both magnetization and polarization could independently encode information in a single multiferroic bit. Four-state memory has recently been demonstrated⁹, but in practice it is likely that the two order parameters are coupled^{10,11}. Coupling could in principle permit data to be written electrically and read magnetically. This is attractive, given that it would exploit the best

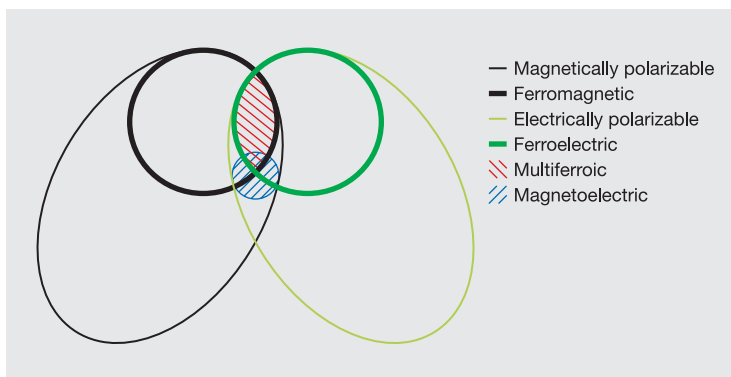


Figure 1 | The relationship between multiferroic and magnetoelectric materials. Ferromagnets (ferroelectrics) form a subset of magnetically (electrically) polarizable materials such as paramagnets and antiferromagnets (paraelectrics and antiferroelectrics). The intersection (red hatching) represents materials that are multiferroic. Magnetoelectric

coupling (blue hatching) is an independent phenomenon that can, but need not, arise in any of the materials that are both magnetically and electrically polarizable. In practice, it is likely to arise in all such materials, either directly or via strain.

¹Department of Materials Science, University of Cambridge, Pembroke Street, Cambridge CB2 3QZ, UK. ²Centre for Ferroics, Earth Sciences Department, University of Cambridge, Downing Street, Cambridge CB2 3EQ, UK.

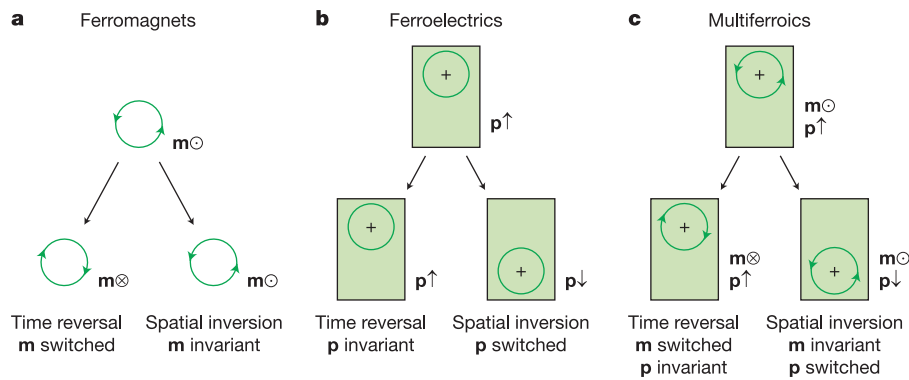


Figure 2 | Time-reversal and spatial-inversion symmetry in ferroics.
a, Ferromagnets. The local magnetic moment **m** may be represented classically by a charge that dynamically traces an orbit, as indicated by the arrowheads. A spatial inversion produces no change, but time reversal switches the orbit and thus **m**. **b**, Ferroelectrics. The local dipole moment **p**

may be represented by a positive point charge that lies asymmetrically within a crystallographic unit cell that has no net charge. There is no net time dependence, but spatial inversion reverses **p**. **c**, Multiferroics that are both ferromagnetic and ferroelectric possess neither symmetry.

aspects of ferroelectric random access memory (FeRAM) and magnetic data storage, while avoiding the problems associated with reading FeRAM and generating the large local magnetic fields needed to write. Unfortunately, significant materials developments will be required to generate magnetoelectric materials that could make a real contribution to the data storage industry. But given the paucity of serious competitors to contemporary memory technologies, the study of novel materials remains important if disruptive technologies are ultimately to emerge. In the shorter term, niche applications are more likely to emerge in strain coupled two-phase systems of the type that we describe later.

The purpose of this Review is to assess the current state of the field, to remind readers of the relevant work performed in the latter half of the twentieth century, and to discuss matters of scientific ‘hygiene’ pertaining to accurate measurements and analyses. For further details we refer the reader to three reviews written at different stages of this re-emerging field^{12–14}.

Magnetoelectric coupling

The magnetoelectric effect in a single-phase crystal is traditionally described^{13,15} in Landau theory by writing the free energy *F* of the system in terms of an applied magnetic field **H** whose *i*th component is denoted *H_i*, and an applied electric field **E** whose *i*th component is denoted *E_i*. Note that this convention is unambiguous in free space, but that *E_i* within a material encodes the resultant field that a test particle would experience. Let us consider a non-ferroic material, where both the temperature-dependent electrical polarization *P_i*(*T*) (μC cm^{−2}) and the magnetization *M_i*(*T*) (μ_B per formula unit, where μ_B is the Bohr magneton) are zero in the absence of applied fields and there is no hysteresis. It may be represented as an infinite, homogeneous and stress-free medium by writing *F* under the Einstein summation convention in S.I. units as:

$$-F(E, H) = \frac{1}{2} \epsilon_0 \epsilon_{ij} E_i E_j + \frac{1}{2} \mu_0 \mu_{ij} H_i H_j + \alpha_{ij} E_i H_j + \frac{\beta_{ijk}}{2} E_i H_j H_k + \frac{\gamma_{ijk}}{2} H_i E_j E_k + \dots$$

(1)

Table 1 Spatial-inversion and time-reversal symmetry in ferroics		
Characteristic symmetry	Spatial-inversion symmetry?	Time-reversal symmetry?
Ferroelastic	Yes	Yes
Ferroelectric	No	Yes
Ferromagnetic	Yes	No
Multiferroic*	No	No

* A multiferroic that is both ferromagnetic and ferroelectric possesses neither symmetry.

The first term on the right hand side describes the contribution resulting from the electrical response to an electric field, where the permittivity of free space is denoted ε₀, and the relative permittivity ε_{*ij*}(*T*) is a second-rank tensor that is typically independent of *E_i* in non-ferroic materials. The second term is the magnetic equivalent of the first term, where μ_{*ij*}(*T*) is the relative permeability and μ₀ is the permeability of free space. The third term describes linear magneto-electric coupling via α_{*ij*}(*T*); the third-rank tensors β_{*ijk*}(*T*) and γ_{*ijk*}(*T*) represent higher-order (quadratic) magnetoelectric coefficients.

In the present scheme, all magnetoelectric coefficients incorporate the field independent material response functions ε_{*ij*}(*T*) and μ_{*ij*}(*T*). The magnetoelectric effects can then easily be established in the form *P_i*(*H_j*) or *M_i*(*E_j*). The former is obtained by differentiating *F* with respect to *E_i*, and then setting *E_i* = 0. A complementary operation involving *H_i* establishes the latter. One obtains:

$$P_i = \alpha_{ij} H_j + \frac{\beta_{ijk}}{2} H_j H_k + \dots$$

(2)

and

$$\mu_0 M_i = \alpha_{ji} E_j + \frac{\gamma_{ijk}}{2} E_j E_k + \dots$$

(3)

In ferroic materials, the above analysis is less rigorous because ε_{*ij*}(*T*) and μ_{*ij*}(*T*) display field hysteresis. Moreover, ferroics are better parameterized in terms of resultant rather than applied fields¹⁶. This is because it is then possible to account for the potentially significant depolarizing/demagnetizing factors in finite media, and also because the coupling constants would then be functions of temperature alone, as in standard Landau theory. In practice, resultant electric and magnetic fields may sometimes be approximated¹⁷ by the polarization and magnetization respectively.

A multiferroic that is ferromagnetic and ferroelectric is liable to display large linear magnetoelectric effects. This follows because ferroelectric and ferromagnetic materials often (but not always) possess a large permittivity and permeability respectively, and α_{*ij*} is bounded by the geometric mean of the diagonalized tensors ε_{*ii*} and μ_{*jj*} such that¹⁸:

$$\alpha_{ij}^2 \leq \epsilon_0 \mu_0 \epsilon_{ii} \mu_{jj}$$

(4)

Equation (4) is obtained from equation (1) by forcing the sum of the first three terms to be greater than zero, that is, ignoring higher-order coupling terms. It represents a stability condition on ε_{*ij*} and μ_{*ij*}, but if the coupling becomes so strong that it drives a phase transition to a more stable state, then α_{*ij*}, ε_{*ij*} and μ_{*ij*} take on new values in the new phase. Note that a large ε_{*ij*} is not a prerequisite for a material to be ferroelectric (or vice versa); and similarly ferromagnets do not necessarily possess large μ_{*ij*}. For example, the ferroelectric KNO₃ possesses a small ε = 25 near its Curie temperature of 120 °C (ref. 19),

Box 1 | Glossary of terms**Ferroics**

Ferroelectric materials possess a spontaneous polarization that is stable and can be switched hysteretically by an applied electric field; **antiferroelectric** materials possess ordered dipole moments that cancel each other completely within each crystallographic unit cell. **Ferromagnetic** materials possess a spontaneous magnetization that is stable and can be switched hysteretically by an applied magnetic field; **antiferromagnetic** materials possess ordered magnetic moments that cancel each other completely within each magnetic unit cell.

Ferroelastic materials display a spontaneous deformation that is stable and can be switched hysteretically by an applied stress.

Ferrotoroidic materials possess a stable and spontaneous order parameter that is taken to be the curl of a magnetization or polarization. By analogy with the above examples, it is anticipated that this order parameter may be switchable. Ferrotoroidic materials have evaded unambiguous observation.

Ferrimagnetic materials differ from antiferromagnets because the magnetic moment cancellation is incomplete in such a way that there is a net magnetization that can be switched by an applied magnetic field.

Order parameter coupling

Magnetoelectric coupling describes the influence of a magnetic (electric) field on the polarization (magnetization) of a material.

Piezoelectricity describes a change in strain as a linear function of applied electric field, or a change in polarization as a linear function of applied stress.

Piezomagnetism describes a change in strain as a linear function of applied magnetic field, or a change in magnetization as a linear function of applied stress.

Electrostriction describes a change in strain as a quadratic function of applied electric field.

Magnetostriction describes a change in strain as a quadratic function of applied magnetic field.

whereas paraelectric SrTiO₃ exhibits $\epsilon > 50,000$ at low temperatures²⁰. Therefore large magnetoelectric couplings need not arise in, or be restricted to, multiferroic materials.

Nonlinear coupling. Most materials have small values of either ϵ_{ij} or μ_{ij} or both, so the linear magnetoelectric effect will also be small, given that permittivity and permeability appear as a product in equation (4). However, no such restriction applies to higher-order couplings, such as those described by β_{ijk} and γ_{ijk} . For example, in some materials terms such as $\beta_{ijk}H_jH_k$ can dominate the linear term $\alpha_{ij}H_j$ in equation (2), as first shown experimentally at low temperatures in the piezoelectric paramagnet NiSO₄·6H₂O (ref. 21). In order to achieve large magnetoelectric effects at room temperature through higher-order terms, we suggest investigating magnetic materials with reduced dimensionality. Indeed, two-dimensional spin order associated with $\beta(T)$ can persist to a temperature T_{2D} that exceeds the temperature T_{3D} at which three-dimensional spin order associated with $\alpha(T)$ is destroyed. This scenario arises²² at low temperature in BaMnF₄.

Indirect coupling. So far, our discussion of linear and higher-order magnetoelectric coupling has ignored the effects of strain. Such effects could be significant or even dominant. For example, the inclusion of piezomagnetism (magnetostriction) would generate cross terms in equation (1) that are proportional to strain and vary linearly (quadratically) with H_i . Analogous expressions would arise from piezoelectricity or electrostriction (see Box 1). Furthermore, mixed terms involving products of strain, H_i and E_j have been predicted²³. In two-phase materials, magnetic and electrical properties are strain-coupled by design in the quest for large magnetoelectric effects. The strength of this indirect coupling is not restricted by equation (4), and enhancements over single-phase systems of several orders of magnitude have been achieved²⁴.

Determination of coupling constants

The magnetoelectric behaviour of a material can only be fully understood if its magnetic point group symmetry is known. This is because the magnetoelectric coefficients α_{ij} , β_{ijk} and γ_{ijk} possess the symmetry of the material^{12,13,15}. For example, α_{ij} can only be non-zero for materials that do not have a centre of symmetry and are time-asymmetric¹². Conversely, information regarding the magnetoelectric coefficients based on electrical²⁵ or optical²⁶ experiments can aid the determination of magnetic point group symmetries.

Below we discuss experimental issues relating to magnetoelectric measurements. The major challenge is to make samples sufficiently insulating to prevent leakage currents contributing to the measured signal—a widespread problem undermining the measurement of ferroelectric polarization loops, as described in refs 16 and 27. Another complication arises if ferroic domains are present, and care should be taken to prepare single-domain states¹¹.

Magnetoelectric coupling can be measured indirectly by simply recording changes in either the magnetization near, say, a ferroelectric transition temperature or the dielectric constant near a magnetic transition temperature. The resulting effects are described using various terms such as ‘magnetocapacitance’ or ‘magnetodielectric response’. Catalan has recently shown²⁸ that the frequently reported effects could misleadingly arise owing to magnetoresistance effects alone, and that the signature of true magnetocapacitance effects is persistence to high frequencies and low loss. However, even true magnetocapacitance measurements do not provide mechanistic insight nor yield coupling constants.

Direct measurements are more challenging. They record either a magnetic response to an applied electric field or an electrical response to an applied magnetic field. The former scenario typically requires electrically addressing the sample in a magnetometer. In the latter scenario, the electrical response can be measured in terms of either current or voltage. The time-integrated current per unit area directly represents the magnetically induced change of polarization in equation (2), that is, $\alpha = \partial P / \partial H$, ignoring higher-order terms. Measurements of voltage, however, yield empirical coupling coefficients commonly also denoted α , which assuming linearity take the form $\partial E / \partial H$.

Single-phase studies

Research into magnetoelectrics and multiferroics took off during the latter half of the twentieth century. In 1957, the linear magnetoelectric coupling coefficient α was predicted to occur in Cr₂O₃ (ref. 29). Then, in the 1960s, α was experimentally observed^{30,31} to be non-zero below the antiferromagnetic Néel temperature of 307 K, near which it peaked to a value of $\alpha = \partial P / \partial H \approx 4.1 \text{ ps m}^{-1}$. This work on Cr₂O₃ and other antiferromagnetic crystals, such as Gd₂CuO₄, Sm₂CuO₄, KNiPO₄, LiCoPO₄ and BiFeO₃, is well summarized in refs 13 and 32.

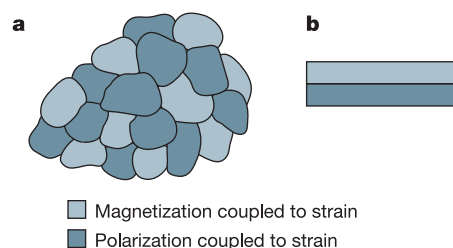


Figure 3 | Strain-mediated magnetoelectric coupling in two-phase systems. Magnetic materials in which the magnetic order parameter couples to strain may be electrically addressed via an intimately connected material that develops a strain in response to an electrical stimulus. Equally, a magnetic stimulus may produce an electrical response. Suitable structures include mixtures of grains (a) and thin-film heterostructures (b). Note that in a but not b the magnetic material must be electrically insulating in order to avoid the possibility of short-circuits.

Magnetoelectric switching. The application potential for multiferroic materials in data storage lies in the possibility of reversing the magnetization by applying an electric field (or vice versa). Magnetoelectric switching was first demonstrated¹¹ 40 years ago in the boracite $\text{Ni}_3\text{B}_7\text{O}_{13}\text{I}$. The weak magnetic and electrical order set in simultaneously¹¹ below 60 K, and a magnetic-field-induced reversal of the magnetization was found to flip the polarization ($0.076 \mu\text{C cm}^{-2}$). This hysteretic response is shown in Fig. 4a. Alternatively, in the paramagnetic ferroelectric $\text{Tb}_2(\text{MoO}_4)_3$, a magnetically induced persistent polarization can arise in large applied magnetic fields⁸ (Fig. 4b). Recently, magnetoelectric switching has been observed in orthorhombic manganites, REMnO_3 or REMn_2O_5 , where RE is a rare earth element. These are antiferromagnets that display improper (weak) ferroelectricity. A small polarization appears at the magnetic Néel temperature (~ 30 K) because the magnetic transition gives rise to crystalline distortions. The polarization of $0.04 \mu\text{C cm}^{-2}$ in TbMn_2O_5 has been magnetically reversed³³, and the polarization of $0.08 \mu\text{C cm}^{-2}$ in TbMnO_3 has been magnetically rotated³ by 90° . Similarly, in the hexaferrite³⁴ $\text{Ba}_{0.5}\text{Sr}_{1.5}\text{Zn}_2\text{Fe}_{12}\text{O}_{22}$, a polarization of $0.015 \mu\text{C cm}^{-2}$ may be magnetically induced and subsequently rotated 360° about the c axis. Unfortunately, these changes in polarization are not persistent, and arise at low temperatures only.

Role of magnetic and crystallographic structure. As magnetoelectric coupling is determined by the structure and magnetic symmetry of a crystal, small modifications might alter, eliminate or allow magnetoelectric effects. Here we discuss key examples.

Spontaneous magnetoelectric coupling in BaMnF_4 permits the ferroelectric order to cant the spins of the two anti-aligned magnetic sub-lattices along the a axis of the crystal, giving a small net ferromagnetic magnetization⁵. The periodicity of this magnetization

is commensurate with the crystal lattice, but the periodicity of the ferroelectricity is incommensurate, and therefore the spatial average of the magnetoelectric coupling is nearly zero and difficult to measure directly. Instead, evidence for coupling comes from changes in the dielectric constant²² observed at magnetic transition temperatures. The magnetization of BaMnF_4 orders two dimensionally below T_{2D} , where a change in the b -axis dielectric constant is observed (Fig. 5). This has been quantified in terms of the magnetic correlation energy^{22,35}. A 1% replacement of Mn with Co destroys the spin canting, and strongly influences the changes in the dielectric constant near the magnetic phase transition, clearly highlighting the key role of magnetic symmetry³⁶.

BiFeO_3 is a commensurate ferroelectric³⁷ and an incommensurate antiferromagnet at room temperature³⁸. The spins are not collinear, but instead take the form of a long-wavelength (62 nm) spiral³⁸. Consequently, the linear magnetoelectric effect also averages to zero here, and indeed only the quadratic effect has been observed³⁹. However, the linear effect may be recovered if the spiral is 'unwound' by applying⁴⁰ large magnetic fields of 20 T, by chemical substitutions⁴¹, or by attempting to introduce thin-film epitaxial constraints⁴². Note that the concept of unwinding has been successfully employed in the case of smectic liquid crystals by confinement within treated glass plates. This surface-stabilized ferroelectric liquid crystal technology (SSFLC) has become the basis for the flat-screen television industry and display screens in cameras.

The linear magnetoelectric effect is also symmetry-forbidden in hexagonal manganites⁴³ of the form REMnO_3 . These materials are ferroelectric, and possess antiferromagnetically aligned Mn^{3+} ions (Néel temperature 70–130 K). RE-site magnetic ordering is seen below ~ 5 K, but large moments may only be achieved in large magnetic fields⁴⁴. Large Néel temperature dielectric anomalies of 42% and 60% are observed in YMnO_3 and HoMnO_3 (ref. 45), respectively. In a landmark experiment, ferroelectric and antiferromagnetic domain walls were seen⁴³ to be coincident in YMnO_3 . Coupling via such walls could then arise because the local magnetization in an antiferromagnetic domain wall either possesses reduced magnetic symmetry⁴⁶, or else interacts with the strain from the coincident ferroelectric wall⁴³. In HoMnO_3 , an electrically driven magnetic phase transition is observed¹⁷. This is a dramatic form of switching that suggests a new approach for magnetoelectric switching.

Circular or toroidal ordering of domains in magnetic or magnetoelectric materials can arise from two unrelated effects. Non-ferroelectric magnetic nanodots typically show vortex (circular) domains⁴⁷ that may be treated as topological defects⁴⁸. In the bulk, ordered magnetic or ferroelectric moments are usually collinear, but toroidal magnetoelectric phenomena have been proposed in ferroelectric ferromagnetic systems^{49,50}. Experimental searches for toroidal ordering⁵¹ became unpopular following irreproducible reports⁵² in

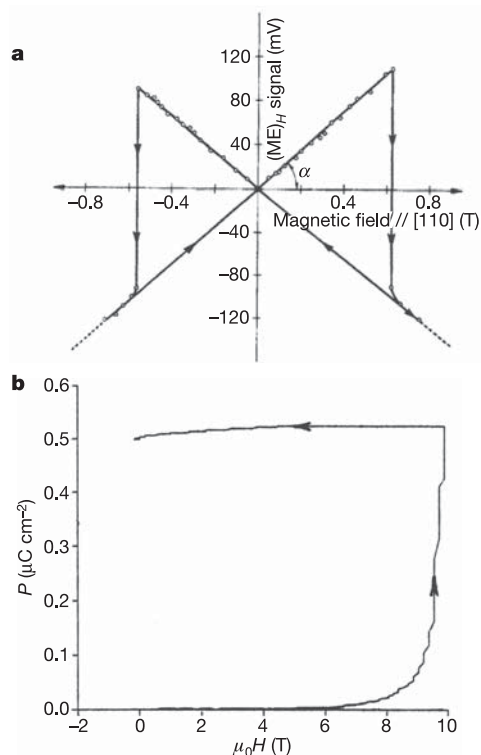


Figure 4 | Examples of magnetoelectric coupling. **a**, $\text{Ni}_3\text{B}_7\text{O}_{13}\text{I}$ at 46 K (after ref. 11, with permission). The magnetoelectric (ME)_H voltage signal, and therefore the polarization via equation (2), may be reversed in magnetic fields ($-0.6 \text{ T} < \mu_0 H < 0.6 \text{ T}$) that are much lower than the switching fields of recently studied materials^{5,55}. **b**, $\text{Tb}_2(\text{MoO}_4)_3$ at 78 K (after ref. 9, with permission). A persistent polarization (P) of $\sim 0.5 \mu\text{C cm}^{-2}$ develops after exposure to magnetic fields in excess of 8 T.

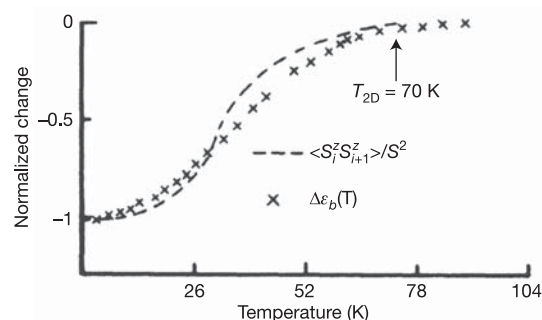


Figure 5 | Link between the magnetic and electrical properties of BaMnF_4 . The temperature dependence of the (7%) change in b -axis dielectric anomaly $\Delta\epsilon_b$ at $T_{2D} = 70$ K closely follows the temperature dependence of the normalized magnetic correlation function $\langle S_i^z S_{i+1}^z \rangle / S^2$. Here S_i^z is the z component of spin S at site i . (After ref. 22, with permission.)

Table 2 | Magnetoelectric coupling constants in two-phase systems

Morphology	Materials	Coupling constant ($\text{mV cm}^{-1} \text{Oe}^{-1}$)	Ref.
Composite	BaTiO ₃ and CoFe ₂ O ₄	50	69
Composite	Terfenol-D and PZT in polymer matrix	42	70
Laminated composites	Terfenol-D in polymer matrix/PZT in polymer matrix	3,000	71
Laminate	Terfenol-D/PZT	4,800	24
Laminate	La _{0.7} Sr _{0.3} MnO ₃ /PZT	60	72
Laminate	NiFe ₂ O ₄ /PZT	1,400	72

PZT (Pb(Zr,Ti)O₃) and BaTiO₃ are piezoelectric, and terfenol-D (Tb_xDy_{1-x}Fe₂), the manganite and the ferrites are magnetostrictive.

1974, but there is now fresh hope that its existence will be experimentally established^{53,54}. The symmetry requirements for toroidal ordering and their relationship with magnetoelectric coupling are reviewed in refs 13 and 14. The area is not well explored, and the possibility of large magnetoelectric coupling exists; however, other possible phenomena, such as ferromagnetism inside antiferromagnetic domain walls, complicate measurements.

Ferromagnetic ferroelectric multiferroics. In 1958, Smolensky's group studied weakly ferromagnetic mixed perovskites, such as $(1-x)\text{Pb}(\text{Fe}_{2/3}\text{W}_{1/3})\text{O}_3-x\text{Pb}(\text{Mg}_{1/2}\text{W}_{1/2})\text{O}_3$ (ref. 55). These were doped to increase resistivity, but the boracite discussed above was the first highly insulating ferromagnetic ferroelectric to be studied¹¹.

The perovskite BiMnO₃ is a promising low-temperature multiferroic that has been considered for quite some time^{56,57}. Ferromagnetic ordering below 105 K is attributed to orbital ordering of the Mn³⁺ ions ($3d^4$) and a large magnetization of $3.6 \mu_B$ per formula unit has been measured for polycrystalline samples⁵⁷. These samples tend to be electrically conducting, restricting evidence of ferroelectricity to low temperatures and precluding definitive measurements of polarization. Multiferroic behaviour was seen⁵⁸ at ~ 80 K ($\sim 1 \mu_B$ per Mn, $<0.15 \mu\text{C cm}^{-2}$), and a peak change of -0.6% in the dielectric constant in 8 T was recorded at T_C (ref. 59). This inspires investigations into less-conducting samples. We have recently shown⁶⁰ that epitaxial films can be highly insulating, even at room temperature ($10^7 \Omega \text{ cm}$).

The spinel CdCr₂S₄ is reported to be a low-temperature multiferroic. In 1965 it was known⁶¹ to be ferromagnetic ($6 \mu_B$ per formula unit) and insulating below 97 K, but only recently was it reported to display⁶² relaxor ferroelectricity ($0.5 \mu\text{C cm}^{-2}$) below 135 K. A colossal magnetocapacitance of 450% in 5 T was also reported, but magnetoresistance effects²⁸ could be responsible.

Reports of room temperature multiferroic materials have centred on BiFeO₃ and its derivatives. Promising multiferroic behaviour ($1 \mu_B$ per formula unit, $\sim 50\text{--}60 \mu\text{C cm}^{-2}$) was reported in thin epitaxial films⁶³, triggering widespread activity. However, subsequent works^{64–66} did not reproduce these findings. The large magnetization only arises in deoxygenated films that are too electrically conducting for valid ferroelectric measurements to be made⁶⁴. Moreover, the presence of impurity phases could play a role^{65,66}. Preliminary reports of multiferroicity in chemically doped BiFeO₃ (ref. 67) and PbTiO₃ (ref. 68) remain unconfirmed.

Two-phase systems

An alternative strategy for engineering enhanced magnetoelectric effects is to introduce indirect coupling, via strain⁶⁹, between two materials such as a ferromagnet and a ferroelectric. Each phase may then be independently optimized for room temperature performance, and the coupling limit of equation (4) is lifted. Strain coupling requires intimate contact (Fig. 3) between a piezomagnetic (or magnetostrictive) material and a piezoelectric (or electrostrictive) material. This can be achieved in the form of composites^{69,70}, laminates^{24,71,72} or epitaxial multilayers⁷³ (Table 2). The coupling constant depends on the frequency of the a.c. applied magnetic field⁷⁴, and such multiferroic structures could thus find applications in, for example, microwave frequency transducers.

Epitaxial thin-film heterostructures could permit precise magne-

toelectric studies, because crystallographic orientation, layer thickness and interfacial roughness may be controlled accurately, but direct measurements of α in epitaxial systems have not been forthcoming. However, ferroelectric layers can generate strains of the order of 1% in magnetic epilayers owing to structural phase transitions. For example, the tetragonal to monoclinic structural phase transition in a BaTiO₃ substrate at 278 K produces⁷³ a 70% change in the magnetization of an epitaxial film of the ferromagnetic manganite La_{0.67}Sr_{0.33}MnO₃ (Fig. 6). Alternatively, one may attempt to alter the magnetic structure of a film by applying a voltage to the underlying piezoelectric material^{75,76,77}. Promising results⁷⁸ were found for a thin film heterostructure of CoPd and Pb(Zr,Ti)O₃ (PZT), where the application of an electric field to the PZT layer rotated the magnetization of the CoPd film by 90°.

The ferromagnetic and ferroelectric phases may be distributed laterally in a film while preserving an epitaxial relationship with one another and the substrate. This has been achieved for nanopillars of CoFe₂O₄ in a BaTiO₃ matrix, grown on a SrRuO₃ electrode with a SrTiO₃ substrate. However, the observed change in magnetization of the CoFe₂O₄ pillars at the ferroelectric Curie temperature was⁷⁹ just 5%, possibly due to either clamping from the underlying epitaxial structure which is not piezoelectric, or electric field effects associated with the ferroelectric. Nevertheless, when the matrix was changed to BiFeO₃, an electrically induced magnetization reversal in the CoFe₂O₄ nanopillars was reported⁸⁰.

Devices. Ferroelectrics may be used to address magnetic materials in devices for two reasons that in practice are not easy to separate^{81–83}. First, their superlative piezoelectric properties permit them to strain intimately connected layers as discussed above. Second, the large

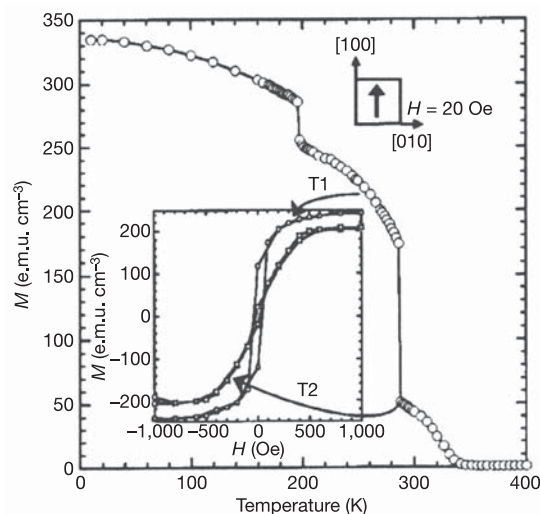


Figure 6 | Magnetoelectric coupling in two-phase systems. Main figure, the magnetization M of an epitaxial La_{0.67}Sr_{0.33}MnO₃ film measured in a small applied field H (top inset) shows sharp changes due to structural phase transitions in the underlying BaTiO₃ substrate. Lower inset, the hard and easy axes are reversed between temperatures T1 and T2. (After ref. 73, with permission.)

polarization can be used in a field-effect transistor geometry to influence the charge density in a magnetic channel.

Various other two-phase magnetoelectric devices that have been explored include a heterostructure comprising PZT and a magnetic garnet between crossed polarizers, where it is possible to electrically influence the Faraday rotation in the garnet and thus control the optical transmission of the device⁸⁴. Exchange bias in $\text{Cr}_2\text{O}_3/(\text{Co/Pt})_3$ may be electrically reversed but requires thermal cycling⁸⁵, whereas exchange bias in YMnO_3 /permalloy heterostructures can be electrically tuned directly⁸⁶. Also, tunable microwave devices with superconductor/ferroelectric/ferromagnetic multilayers have been proposed⁸⁷. In other devices, strain-coupled magnetostrictive and piezoelectric layers can lead to voltage gain⁸⁸, and the detection of magnetic fields⁸⁹. The sensor devices seem particularly promising compared to existing superconducting quantum interference device (SQUID) technology because not only would they be cheaper and simpler, but also they can operate at room temperature.

Outlook

The flurry of activity during the past six years has not yet produced a great leap forward for multiferroics and magnetoelectrics, for two reasons. First, although new materials such as the manganite TbMnO_3 have come to light³, they do not necessarily compare favourably with the boracite¹¹ $\text{Ni}_3\text{B}_7\text{O}_{13}\text{I}$, which was studied almost three decades before Schmid coined the term⁷ 'multiferroic'. The boracite possesses smaller switching fields and higher ordering temperatures than the manganite. Second, most of the modern systems under study remain to be formally analysed in terms of magnetic point group symmetries, which are vital for the analysis and prediction of magnetoelectric effects.

There are nevertheless plenty of reasons to be optimistic about the health of the field in both the short and the long term. A unified quantum mechanical theoretical approach may be possible in the future because magnetism is necessarily a quantum mechanical phenomenon, and modern theories of ferroelectricity calculate the polarization as a quantum mechanical observable⁹⁰ rather than using the semi-classical ball-and-stick model that has hitherto been so successful. Therefore *ab initio* modelling could in future show the way forward. In fact, there is plenty of scope for development, given that the ferroelectric moment cannot necessarily be determined by the rigid displacement of an ion. This is the case whenever electronic orbitals are not tied rigidly to ions^{91–93}.

Several strategies for achieving properties of interest also give rise to optimism. For example, ferrimagnetic materials are more likely to be electrically insulating than ferromagnets, and indeed the classic lodestone material Fe_3O_4 shows a magnetically addressable⁹⁴ polarization⁹⁵ of $5\ \mu\text{C cm}^{-2}$ below the Verwey transition at 120 K. Alternatively, non-centrosymmetric structures may be achieved in thin films by exploiting epitaxial strain, defects or composition gradients, for example, in garnets⁹⁶. Epitaxial multilayers offer particularly close control of the strain coupling discussed earlier⁹⁷.

Magnetic switching by a voltage has the chance to compete with magnetic switching driven by light⁹⁸ or spin-polarized currents⁹⁹. Devices that would benefit from single-phase magnetoelectric materials (that are ferromagnetic) include, for instance, spin filters¹⁰⁰ for data storage and spatial light modulators⁸⁴ for optical switching.

We close by highlighting several issues that focus upon the scientific and technological promise of multiferroic and magnetoelectric materials. First, it remains to be seen whether there exists a ferromagnetic ferroelectric multiferroic in which the two order parameters are both either strong at any temperature, or non-zero at room temperature. Second, reports of electrically induced magnetic reversal remain rare. Third, a deeper understanding of single-phase magnetoelectric materials would be enhanced by establishing whether couplings are direct or strain-mediated. Fourth, the symmetry of magnetoelectric systems under study can yield valuable information. Fifth, we stress that ferroelectricity arises in a wide range of materials systems

including fluorides, and the search should not be limited to oxides. Sixth, the acquisition of ferroelectric data differs from magnetic measurements in that it involves a complete system with contacts and leads; the results are prone to misinterpretation^{16,27} if sample resistivities stray far below $10^7\ \Omega\text{ cm}$. Seventh, we reiterate that the nomenclature of the field is confused. Consequently 'the field' is an imperfectly defined entity, as reflected in the composite title of this Review. Finally, we suggest that the complex materials discussed here are in their infancy. Therefore the current scientific interest may one day lead to developments that contribute to technology.

- Valasek, J. Piezoelectric and allied phenomena in Rochelle salt. *Phys. Rev.* **15**, 537–538 (1920).
- Mathur, N. D. *et al.* Magnetically mediated superconductivity in heavy fermion compounds. *Nature* **394**, 39–43 (1998).
- Kimura, T. *et al.* Magnetic control of ferroelectric polarization. *Nature* **426**, 55–58 (2003).
- Scott, J. F. Phase transitions in BaMnF_4 . *Rep. Prog. Phys.* **42**, 1055–1084 (1979).
- Fox, D. L. & Scott, J. F. Ferroelectrically induced ferromagnetism. *J. Phys. C* **10**, L329–L331 (1977).
- Hill, N. A. Why are there so few magnetic ferroelectrics? *J. Phys. Chem. B* **104**, 6694–6709 (2000).
- Schmid, H. Multi-ferroic magnetoelectrics. *Ferroelectrics* **162**, 665–685 (1994).
- Ponomarev, B. K. *et al.* Magnetoelectric properties of some rare earth molybdates. *Ferroelectrics* **161**, 43–48 (1994).
- Gajek, M. *et al.* Multiferroic tunnel junctions. Preprint at (<http://arxiv.org/cond-mat/0606444>) (2006).
- Schmid, H. On a magnetoelectric classification of materials. *Int. J. Magn.* **4**, 337–361 (1973).
- Ascher, E., Rieder, H., Schmid, H. & Stössel, H. Some properties of ferromagnetoelectric nickel-iodine boracite, $\text{Ni}_3\text{B}_7\text{O}_{13}\text{I}$. *J. Appl. Phys.* **37**, 1404–1405 (1966).
- O'Dell, T. H. *The Electrodynamics of Magneto-electric Media* (North-Holland, Amsterdam, 1970).
- Schmid, H. Introduction to the proceedings of the 2nd international conference on magnetoelectric interaction phenomena in crystals, MEIPIC-2. *Ferroelectrics* **161**, 1–28 (1994).
- Fiebig, M. Revival of the magnetoelectric effect. *J. Phys. D* **38**, R123–R152 (2005).
- Rivera, J.-P. On definition, units, measurements, tensor forms of the linear magnetoelectric effect and on a new dynamic method applied to Cr-Cl boracite. *Ferroelectrics* **161**, 165–180 (1994).
- Lines, M. E. & Glass, A. M. *Principles and Applications of Ferroelectrics and Related Materials* (Clarendon Press, Oxford, 1977).
- Lottermoser, T. *et al.* Magnetic phase control by an electric field. *Nature* **430**, 541–544 (2004).
- Brown, W. F. Jr, Hornreich, R. M. & Shtrikman, S. Upper bound on the magnetoelectric susceptibility. *Phys. Rev.* **168**, 574–577 (1968).
- Chen, A. & Chernow, F. Nature of ferroelectricity in KNO_3 . *Phys. Rev.* **154**, 493–505 (1967).
- Saifi, M. A. & Cross, L. E. Dielectric properties of strontium titanate at low temperatures. *Phys. Rev. B* **2**, 677–684 (1970).
- Hou, S. L. & Bloembergen, N. Paramagnetoelectric effects in $\text{NiSO}_4 \cdot 6\text{H}_2\text{O}$. *Phys. Rev.* **138**, A1218–A1226 (1965).
- Scott, J. F. Mechanisms of dielectric anomalies in BaMnF_4 . *Phys. Rev. B* **16**, 2329–2331 (1977).
- Grimmer, H. The piezomagnetoelectric effect. *Acta Crystallogr. A* **48**, 266–271 (1992).
- Ryu, J., Vásquez Carazo, A., Uchino, K. & Kim, H.-E. Magnetoelectric properties in piezoelectric and magnetostrictive laminate composites. *Jpn. J. Appl. Phys.* **40**, 4948–4951 (2001).
- Sciau, Ph., Clin, M., Rivera, J.-P. & Schmid, H. Magnetoelectric measurements on BaMnF_4 . *Ferroelectrics* **105**, 201–206 (1990).
- Fiebig, M. *et al.* Determination of the magnetic symmetry of hexagonal manganites by second harmonic generation. *Phys. Rev. Lett.* **84**, 5620–5623 (2000).
- Dawber, M., Rabe, K. & Scott, J. F. Physics of ferroelectric thin film oxides. *Rev. Mod. Phys.* **77**, 1083–1130 (2005).
- Catalan, G. Magnetocapacitance without magnetoelectric coupling. *Appl. Phys. Lett.* **88**, 102902 (2006).
- Dzyaloshinskii, I. E. On the magneto-electrical effects in antiferromagnets. *Zh. Eksp. Teor. Fiz.* **37**, 881–882 [*Sov. Phys. JETP* **10**, 628–629] (1959).
- Astrov, D. N. The magnetoelectric effect in antiferromagnetics. *Zh. Eksp. Teor. Fiz.* **38**, 984–985 [*Sov. Phys. JETP* **11**, 708–709] (1960).
- Folen, V. J., Rado, G. T. & Stalder, E. W. Anisotropy of the magnetoelectric effect in Cr_2O_3 . *Phys. Rev. Lett.* **6**, 607–608 (1961).
- Bichurin, M. Short introduction to the proceedings of the 3rd International Conference on Magnetoelectric Interaction Phenomena in Crystals, MEIPIC-3. *Ferroelectrics* **204**, XVII–XX (1997).
- Hur, N. *et al.* Electric polarization reversal and memory in a multiferroic material induced by magnetic fields. *Nature* **429**, 392–395 (2004).

34. Kimura, T., Lawes, G. & Ramirez, A. P. Electric polarization rotation in a hexaferrite with long-wavelength magnetic structures. *Phys. Rev. Lett.* **94**, 137201 (2005).
35. de Jongh, L. J. & Miedema, A. R. Experiments on simple magnetic model systems. *Adv. Phys.* **23**, 1–260 (1974).
36. Fox, D. L., Tilley, D. R., Scott, J. F. & Guggenheim, H. J. Magnetoelectric phenomena in BaMnF_4 and $\text{BaMn}_{0.99}\text{Co}_{0.01}\text{F}_4$. *Phys. Rev. B* **21**, 2926–2936 (1980).
37. Teage, J. R., Gerson, R. & James, W. J. Dielectric hysteresis in single crystal BiFeO_3 . *Solid State Commun.* **8**, 1073–1074 (1970).
38. Sosnovska, I., Peterlin-Neumaier, T. & Steichele, E. Spiral magnetic ordering in bismuth ferrite. *J. Phys. C* **15**, 4835–4846 (1982).
39. Tabarez-Muñoz, C. *et al.* Measurement of the quadratic magnetoelectric effect on single crystalline BiFeO_3 . *Jpn. J. Appl. Phys.* **24**, 1051–1053 (1985).
40. Popov, Y. F., Kadamtsseva, A. M., Vorob'ev, G. P. & Zvezdin, A. K. Discovery of the linear magnetoelectric effect in magnetic ferroelectric BiFeO_3 in a strong magnetic field. *Ferroelectrics* **162**, 135–140 (1994).
41. Murashov, V. A. *et al.* Magnetoelectric $(\text{Bi},\text{Ln})\text{FeO}_3$ compounds: crystal growth, structure and properties. *Ferroelectrics* **162**, 11–21 (1994).
42. Bai, F. *et al.* Destruction of spin cycloid in $(111)_c$ -oriented BiFeO_3 thin films by epitaxial constraint: Enhanced polarization and release of latent magnetization. *Appl. Phys. Lett.* **86**, 032511 (2005).
43. Fiebig, M. *et al.* Observation of coupled magnetic and electric domains. *Nature* **419**, 818–820 (2002).
44. Sugie, H., Iwata, N. & Kohn, K. Magnetic ordering of rare earth ions and magnetic-electric interaction of hexagonal RMnO_3 ($R = \text{Ho}, \text{Er}, \text{Yb}$ or Lu). *J. Phys. Soc. Jpn* **71**, 1558–1564 (2002).
45. Lorenz, B., Wang, Y. Q., Sun, Y. Y. & Chu, C. W. Large magnetodielectric effects in orthorhombic HoMnO_3 and YMnO_3 . *Phys. Rev. B* **70**, 212412 (2004).
46. Lottermoser, T. & Fiebig, M. Magnetoelectric behavior of domain walls in multiferroic HoMnO_3 . *Phys. Rev. B* **70**, 220407(R) (2004).
47. Kittel, C. Theory of the structure of ferromagnetic domains in films and small particles. *Phys. Rev.* **70**, 965–971 (1946).
48. Mermin, N. D. The topological theory of defects in ordered media. *Rev. Mod. Phys.* **51**, 591–648 (1979).
49. Ginzburg, V. L. *et al.* On the problem of superdiamagnetism. *Solid State Commun.* **50**, 339–343 (1984).
50. Gorbatshevich, A. A. & Kopaev, Y. V. Toroidal order in crystals. *Ferroelectrics* **161**, 321–334 (1994).
51. Schmid, H. On ferrotoroidics and electrotoroidic, magnetotoroidic and piezotoroidic effects. *Ferroelectrics* **252**, 253–268 (2001).
52. Zheludev, I. S. *et al.* Magnetic properties of nickel iodine boracite. *JETP Lett.* **20**, 129–130 (1974).
53. Naumov, I. V., Bellaiche, L. & Fu, H. Unusual phase transitions in ferroelectric nanodisks and nanorods. *Nature* **432**, 737–740 (2004).
54. Scott, J. F. Ferroelectrics: Novel geometric ordering of ferroelectricity. *Nature Mater.* **4**, 13–14 (2005).
55. Smolensky, G. A., Isupov, V. A. & Agronovskaya, A. I. New ferroelectrics of complex composition of the type $\text{A}_2^{2+}(\text{B}_1^{3+}\text{B}_2^{5+})\text{O}_6$. *Sov. Phys. Solid State* **1**, 150–151 (1959).
56. Sugawara, F., Iida, S., Syono, Y. & Akimoto, S. Magnetic properties and crystal distortions of BiMnO_3 and BiCrO_3 . *J. Phys. Soc. Jpn* **25**, 1553–1558 (1968).
57. Chiba, H., Atou, T. & Syono, Y. Magnetic and electrical properties of $\text{Bi}_{1-x}\text{Sr}_x\text{MnO}_3$. *J. Solid State Chem.* **132**, 139–143 (1997).
58. Moreira dos Santos, A. *et al.* Evidence for the likely occurrence of magnetoferroelectricity in the simple perovskite, BiMnO_3 . *Solid State Commun.* **122**, 49–52 (2002).
59. Kimura, T. *et al.* Magnetocapacitance effect in multiferroic BiMnO_3 . *Phys. Rev. B* **67**, R180401 (2003).
60. Eerenstein, W., Morrison, F. D., Scott, J. F. & Mathur, N. D. Growth of highly resistive BiMnO_3 films. *Appl. Phys. Lett.* **87**, 101906 (2005).
61. Baltzer, P. K., Lehmann, H. W. & Robbins, M. Insulating ferromagnetic spinels. *Phys. Rev. Lett.* **15**, 493–495 (1965).
62. Hemberger, J. *et al.* Relaxor ferroelectricity and colossal magnetocapacitive coupling in ferromagnetic CdCr_2S_4 . *Nature* **434**, 364–367 (2005).
63. Wang, J. *et al.* Epitaxial BiFeO_3 multiferroic thin film heterostructures. *Science* **299**, 1719–1722 (2003).
64. Eerenstein, W. *et al.* Comment on “Epitaxial BiFeO_3 multiferroic thin film heterostructures”. *Science* **419**, 1203a (2005).
65. Wang, J. *et al.* Response to Comment on “Epitaxial BiFeO_3 multiferroic thin film heterostructures”. *Science* **419**, 1203b (2005).
66. Béa, H. *et al.* Influence of parasitic phases on the properties of BiFeO_3 epitaxial thin films. *Appl. Phys. Lett.* **87**, 072508 (2005).
67. Palkar, V. R., Kundaliya, D. C., Malik, S. K. & Bhattacharya, S. Magnetoelectricity at room temperature in the $\text{Bi}_{0.9-x}\text{Tb}_x\text{La}_{0.1}\text{FeO}_3$ system. *Phys. Rev. B* **69**, 212102 (2004).
68. Palkar, V. R. & Malik, S. K. Observation of magnetoelectric behavior at room temperature in $\text{Pb}(\text{Fe}_2\text{Ti}_{1-x})\text{O}_3$. *Solid State Commun.* **134**, 783–786 (2005).
69. van Run, A. M. J. G., Terrell, D. R. & Scholing, J. H. An *in situ* grown eutectic magnetoelectric composite material. *J. Mater. Sci.* **9**, 1710–1714 (1974).
70. Nan, C.-W. *et al.* A three-phase magnetoelectric composite of piezoelectric ceramics, rare-earth iron alloys, and polymer. *Appl. Phys. Lett.* **81**, 3831–3833 (2002).
71. Cai, N., Nan, C.-W., Zhai, J. & Lin, Y. Large high-frequency magnetoelectric response in laminated composites of piezoelectric ceramics, rare-earth iron alloys and polymer. *Appl. Phys. Lett.* **84**, 3516–3519 (2004).
72. Srinivasan, G. *et al.* Magnetoelectric bilayer and multilayer structures of magnetostrictive and piezoelectric oxides. *Phys. Rev. B* **65**, 134402 (2002).
73. Lee, M. K. *et al.* Strain modification of epitaxial perovskite oxide thin films using structural transitions of ferroelectric BaTiO_3 substrate. *Appl. Phys. Lett.* **77**, 3547–3549 (2000).
74. Bichurin, M. I. *et al.* Resonance magnetoelectric effects in layered magnetostrictive-piezoelectric composites. *Phys. Rev. B* **68**, 132408 (2003).
75. Schröder, K. Stress operated random access, high speed magnetic memory. *J. Appl. Phys.* **53**, 2759–2761 (1982).
76. Novosad, V. *et al.* Novel magnetostrictive memory device. *J. Appl. Phys.* **87**, 6400–6402 (2000).
77. Iwasaki, Y. Stress-driven magnetization reversal in magnetostrictive films with in-plane magnetocrystalline anisotropy. *J. Magn. Magn. Mater.* **240**, 395–397 (2002).
78. Kim, S.-K. *et al.* Voltage control of a magnetization easy axis in piezoelectric/ferromagnetic hybrid films. *J. Magn. Magn. Mater.* **267**, 127–132 (2003).
79. Zheng, H. *et al.* Multiferroic $\text{BaTiO}_3\text{-CoFe}_2\text{O}_4$ nanostructures. *Science* **303**, 661–663 (2004).
80. Zavaliche, F. *et al.* Electric field-induced magnetization switching in epitaxial columnar nanostructures. *Nano Lett.* **5**, 1793–1796 (2005).
81. Mathews, S., Ramesh, R., Venkatesan, T. & Benedetto, J. Ferroelectric field effect transistor based on epitaxial perovskite heterostructures. *Science* **276**, 238–240 (1997).
82. Wu, T. *et al.* Electroresistance and electronic phase separation in mixed-valent manganites. *Phys. Rev. Lett.* **86**, 5998–6001 (2001).
83. Tabata, H. & Kawai, T. Novel electronic properties of ferroelectric/ferromagnetic heterostructures. *IEICI Tran. Electron.* **E80-C**, 918–922 (1997).
84. Inoue, M. Magnetophotonic crystals. *Mater. Res. Soc. Symp. Proc.* **834**, J1.1–J1.19 (2005).
85. Borisov, P. *et al.* Magnetoelectric switching of exchange bias. *Phys. Rev. Lett.* **94**, 117203 (2005).
86. Laukmin, V. *et al.* Electric-field control of exchange bias in multiferroic epitaxial heterostructures. Preprint at (<http://arxiv.org/cond-mat/0607381>) (2006).
87. Hontsu, S. *et al.* Preparation of all-oxide ferromagnetic/ferroelectric/superconducting heterostructures for advanced microwave applications. *Supercond. Sci. Technol.* **12**, 836–839 (1999).
88. Dong, S. *et al.* A strong magnetoelectric voltage gain effect in magnetostrictive-piezoelectric composite. *Appl. Phys. Lett.* **85**, 3534–3536 (2004).
89. Dong, S., Li, J. F. & Viehland, D. Ultrahigh magnetic field sensitivity in laminates of Terfenol-D and $\text{Pb}(\text{Mg}_{1/3}\text{Nb}_{2/3})\text{O}_3\text{-PbTiO}_3$ crystals. *Appl. Phys. Lett.* **83**, 2265–2267 (2003).
90. Kingsmith, R. D. & Vanderbilt, D. Theory of polarization in crystalline solids. *Phys. Rev. B* **47**, 1651–1654 (1993).
91. Jaynes, E. T. *Ferroelectricity* (Princeton Univ. Press, Princeton, 1953).
92. Littlewood, P. B. & Heine, V. The infrared effective charge in IV–VI compounds: I. A simple one-dimensional model. *J. Phys. C* **12**, 4431–4439 (1979).
93. Efremov, D. V., Van den Brink, J. & Khomskii, D. I. Bond- versus site-ordering and possible ferroelectricity in manganites. *Nature Mater.* **3**, 853–856 (2004).
94. Rado, G. T. & Ferrari, J. M. Linear and bilinear magnetoelectric effects in magnetically biased magnetite (Fe_3O_4). *Phys. Rev. B* **15**, 290–297 (1977).
95. Kato, K., Iida, S., Yanai, K. & Mizushima, K. Ferrimagnetic ferroelectricity of Fe_3O_4 . *J. Magn. Magn. Mater.* **31–34**, 783–784 (1983).
96. Krichetsov, B. B. *et al.* Linear magnetoelectric effect in magnetic garnet thin films. *Ferroelectrics* **161**, 65–71 (1994).
97. Yamada, Y., Kawasaki, M., Ogawa, Y. & Tokura, Y. Perovskite oxide tricolor superlattices with artificially broken inversion symmetry by interface effects. *Appl. Phys. Lett.* **81**, 4793–4795 (2002).
98. Kimel, A. V. *et al.* Ultrafast non-thermal control of magnetization by instantaneous photomagnetic pulses. *Nature* **435**, 655–657 (2005).
99. Sun, J. Z. Current-driven magnetic switching in manganite trilayer junctions. *J. Magn. Magn. Mater.* **202**, 157–162 (1999).
100. Gajek, M. *et al.* Spin filtering through ferromagnetic BiMnO_3 tunnel barriers. *Phys. Rev. B* **72**, 020406(R) (2005).

Acknowledgements We thank P. B. Littlewood, M. Fiebig and A. D. Kent for discussions and S. Celotto for assistance with reproducing figures. This work was supported by an EU Marie Curie Fellowship (W.E.), The Royal Society (N.D.M.) and the UK EPSRC.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence should be addressed to N.D.M. (ndm12@cam.ac.uk).

ARTICLES

Driven coherent oscillations of a single electron spin in a quantum dot

F. H. L. Koppens¹, C. Buizert¹, K. J. Tielrooij¹, I. T. Vink¹, K. C. Nowack¹, T. Meunier¹, L. P. Kouwenhoven¹ & L. M. K. Vandersypen¹

The ability to control the quantum state of a single electron spin in a quantum dot is at the heart of recent developments towards a scalable spin-based quantum computer. In combination with the recently demonstrated controlled exchange gate between two neighbouring spins, driven coherent single spin rotations would permit universal quantum operations. Here, we report the experimental realization of single electron spin rotations in a double quantum dot. First, we apply a continuous-wave oscillating magnetic field, generated on-chip, and observe electron spin resonance in spin-dependent transport measurements through the two dots. Next, we coherently control the quantum state of the electron spin by applying short bursts of the oscillating magnetic field and observe about eight oscillations of the spin state (so-called Rabi oscillations) during a microsecond burst. These results demonstrate the feasibility of operating single-electron spins in a quantum dot as quantum bits.

The use of quantum mechanical superposition states and entanglement in a computer can theoretically solve important mathematical and physical problems much faster than classical computers^{1,2}. However, the realization of such a quantum computer represents a formidable challenge, because it requires fast and precise control of fragile quantum states. The prospects for accurate quantum control in a scalable system are thus being explored in a rich variety of physical systems, ranging from nuclear magnetic resonance and ion traps to superconducting devices³.

Electron spin states were identified early on as an attractive realization of a quantum bit⁴, because they are relatively robust against decoherence (uncontrolled interactions with the environment). Advances in the field of semiconductor quantum dots have made this system very fruitful as a host for the electron spin. Since Loss and DiVincenzo's proposal⁵ on electron spin qubits in quantum dots in 1998, many of the elements necessary for quantum computation have been realized experimentally. It is now routine to isolate with certainty a single electron in each of two coupled quantum dots^{6–9}. The spin of this electron can be reliably initialized to the ground state, spin-up, via optical pumping¹⁰ or by thermal equilibration at sufficiently low temperatures and strong static magnetic fields (for example, $T = 100$ mK and $B_{\text{ext}} = 1$ T). The spin states are also very long-lived, with relaxation times of the order of milliseconds^{11–13}. Furthermore, a lower bound on the spin coherence time exceeding 1 μ s was established, using spin-echo techniques on a two-electron system¹⁴. These long relaxation and coherence times are possible in part because the magnetic moment of a single electron spin is so weak. On the other hand, this property makes read-out and manipulation of single spins particularly challenging. By combining spin-to-charge conversion with real-time single-charge detection^{15–17}, it has nevertheless been possible to accomplish single-shot read-out of spin states in a quantum dot^{13,18}.

The next major achievement was the observation of the coherent exchange of two electron spins in a double dot system, controlled by fast electrical switching of the tunnel coupling between the two quantum dots¹⁴. Finally, free evolution of a single electron spin about

a static magnetic field (Larmor precession) has been observed, via optical pump-probe experiments^{19,20}. The only missing ingredient for universal quantum computation with spins in dots remained the demonstration of driven coherent spin rotations (Rabi oscillations) of a single electron spin.

The most commonly used technique for inducing spin flips is electron spin resonance (ESR)²¹. ESR is the physical process whereby electron spins are rotated by an oscillating magnetic field B_{ac} (with frequency f_{ac}) that is resonant with the spin precession frequency in an external magnetic field B_{ext} , oriented perpendicularly to B_{ac} ($\hbar f_{\text{ac}} = g\mu_B B_{\text{ext}}$, where μ_B is the Bohr magneton and g the electron spin g -factor). Magnetic resonance of a single electron spin in a solid has been reported in a few specific cases^{22–24}, but has never been realized in semiconductor quantum dots. Detecting ESR in a single quantum dot is conceptually simple²⁵, but experimentally difficult to realize, as it requires a strong, high-frequency magnetic field at low temperature, while accompanying alternating electric fields must be minimized. Alternative schemes for driven rotations of a spin in a dot have been proposed, based on optical excitation²⁶ or electrical control^{27–29}, but this is perhaps even more challenging and has not been accomplished either.

Here, we demonstrate the ability to control the spin state of a single electron confined in a double quantum dot via ESR. In a double dot system, spin-flips can be detected through the transition of an electron from one dot to the other^{30,31} rather than between a dot and a reservoir, as would be the case for a single dot. This has the advantage that there is no need for the electron spin Zeeman splitting (used in a single dot for spin-selective tunnelling) to exceed the temperature of the electron reservoirs (~ 100 mK; the phonon temperature was ~ 40 mK). The experiment can thus be performed at a smaller static magnetic field, and consequently with lower, technically less demanding, excitation frequencies. Furthermore, by applying a large bias voltage across the double dot, the spin detection can be made much less sensitive to electric fields than is possible in the single-dot case (electric fields can cause photon-assisted tunnelling; see Supplementary Discussion). Finally, in a double dot, single-spin

¹Kavli Institute of NanoScience, Delft University of Technology, PO Box 5046, 2600 GA, Delft, The Netherlands.

operations can in future experiments be combined with two-qubit operations to realize universal quantum gates⁵, and with spin read-out to demonstrate entanglement^{32,33}.

Device and ESR detection concept

Two coupled semiconductor quantum dots are defined by surface gates (Fig. 1a) on top of a two-dimensional electron gas. By applying the appropriate negative voltages to the gates the dots can be tuned to the few-electron regime⁸. The oscillating magnetic field that drives the spin transitions is generated by applying a radio-frequency (RF) signal to an on-chip coplanar stripline (CPS) which is terminated in a narrow wire, positioned near the dots and separated from the surface gates by a 100-nm-thick dielectric (Fig. 1b). The current through the wire generates an oscillating magnetic field B_{ac} at the dots, perpendicular to the static external field B_{ext} and slightly stronger in the left dot than in the right dot (see Supplementary Fig. S1).

To detect the ESR-induced spin rotations, we use electrical transport measurements through the two dots in series in the spin blockade regime where current flow depends on the relative spin state of the electrons in the two dots^{30,34}. In brief, the device is operated so that current is blocked owing to spin blockade, but this blockade is lifted if the ESR condition ($hf_{ac} = g\mu_B B_{ext}$) is satisfied.

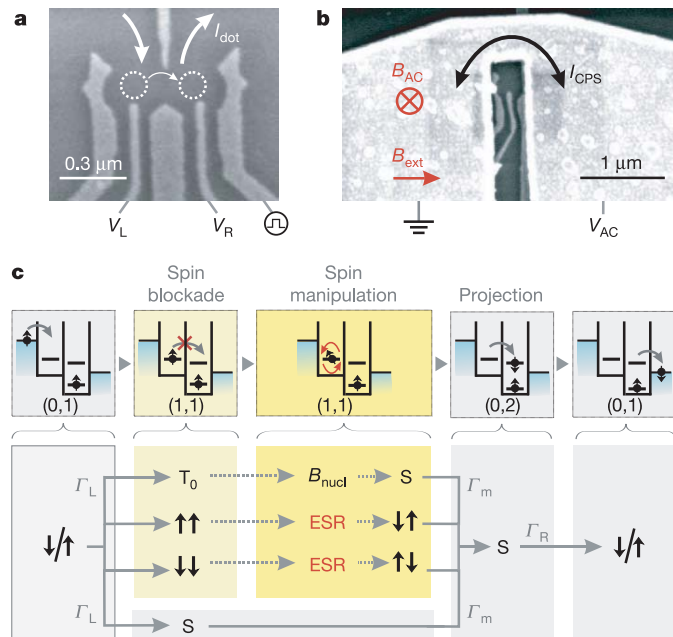


Figure 1 | Device and ESR detection scheme. **a**, Scanning electron microscope (SEM) image of a device with the same gate pattern as used in the experiment. The Ti/Au gates are deposited on top of a GaAs/AlGaAs heterostructure containing a two-dimensional electron gas 90 nm below the surface. White arrows indicate current flow through the two coupled dots (dotted circles). The right side gate is fitted with a homemade bias-tee (rise time 150 ps) to allow fast pulsing of the dot levels. **b**, SEM image of a device similar to the one used in the experiment. The termination of the coplanar stripline is visible on top of the gates. The gold stripline has a thickness of 400 nm and is designed to have a 50 Ω characteristic impedance, Z_0 , up to the shorted termination. It is separated from the gate electrodes by a 100-nm-thick dielectric (Calixerene)³⁰. **c**, Diagrams illustrating the transport cycle in the spin blockade regime. This cycle can be described via the occupations (m,n) of the left and right dots as $(0,1) \rightarrow (1,1) \rightarrow (0,2) \rightarrow (0,1)$. When an electron enters the left dot (with rate Γ_L) starting from $(0,1)$, the two-electron system that is formed can be either a singlet $S(1,1)$ or a triplet $T(1,1)$. From $S(1,1)$, further current flow is possible via a transition to $S(0,2)$ (with rate Γ_m). When the system is in $T(1,1)$, current is blocked unless this state is coupled to $S(1,1)$. For T_0 , this coupling is provided by the inhomogeneous nuclear field ΔB_N . For $T_{\pm}$ or $T_{\mp}$, ESR causes a transition to $\uparrow\downarrow$ or $\downarrow\uparrow$, which contains a $S(1,1)$ component and a T_0 component (which is in turn coupled to $S(1,1)$ by the nuclear field).

This spin blockade regime is accessed by tuning the gate voltages such that one electron always resides in the right dot, and a second electron can tunnel from the left reservoir to the left dot (Fig. 1c and Supplementary Fig. S2). If this electron forms a double-dot singlet state with the electron in the right dot ($S = \uparrow\downarrow - \downarrow\uparrow$; normalization omitted for brevity), it is possible for the left electron to move to the right dot, and then to the right lead (leaving behind an electron in the right dot with spin $\uparrow$ or spin $\downarrow$), since the right dot singlet state is energetically accessible. If, however, the two electrons form a double-dot triplet state, the left electron cannot move to the right dot because the right dot's triplet state is much higher in energy. The electron also cannot move back to the lead and therefore further current flow is blocked as soon as any of the (double-dot) triplet states is formed.

Role of the nuclear spin bath for ESR detection

In fact, the situation is more complex, because each of the two spins experiences a randomly oriented and fluctuating effective nuclear field of ~ 1 –3 mT (refs 35, 36). This nuclear field, B_N , arises from the hyperfine interaction of the electron spins with the Ga and As nuclear spins in the host material, and is in general different in the two dots, with a difference of ΔB_N . At zero external field and for sufficiently small double dot singlet–triplet splitting (see Supplementary Fig. S2d), the inhomogeneous component of the nuclear field causes all three triplet states (T_0 , T_{+} and T_{-}) to be admixed with the singlet S (for example, $T_0 = \uparrow\downarrow + \downarrow\uparrow$ evolves into $S = \uparrow\downarrow - \downarrow\uparrow$ due to $\Delta B_{N,z}$ and $T_{+} = \uparrow\uparrow$ and $T_{-} = \downarrow\downarrow$ evolve into S owing to $\Delta B_{N,x}$). As a result, spin blockade is lifted. For $B_{ext} \gg \sqrt{B_N^2}$, however, the T_{+} and T_{-} states split off in energy, which makes hyperfine-induced admixing between $T_{\pm}$ and S ineffective (T_0 and S remain admixed; see Fig. 2a). Here spin blockade does occur, whenever a state with parallel spins ($\uparrow\uparrow$ or $\downarrow\downarrow$) becomes occupied.

ESR is then detected as follows (see Fig. 1c). An oscillating magnetic field resonant with the Zeeman splitting can flip the spin in the left or the right dot. Starting from $\uparrow\uparrow$ or $\downarrow\downarrow$, the spin state then changes to $\uparrow\downarrow$ (or $\downarrow\uparrow$). If both spins are flipped, transitions occur between $\uparrow\uparrow$ and $\downarrow\downarrow$ via the intermediate state $\frac{1}{\sqrt{2}}(\uparrow\downarrow + \downarrow\uparrow)$. In both cases, states with anti-parallel spins ($S_z = 0$) are created owing to ESR. Expressed in the singlet-triplet measurement basis, $\uparrow\downarrow$ or $\downarrow\uparrow$ is a superposition of the T_0 and S state ($\uparrow\downarrow = T_0 + S$). For the singlet component of this state, the left electron can transition immediately to the right dot and from there to the right lead. The T_0 component first evolves into a singlet due to the nuclear field and then the left electron can move to the right dot as well. Thus whenever the spins are anti-parallel, one electron charge moves through the dots. If such transitions from parallel to anti-parallel spins are induced repeatedly at a sufficiently high rate, a measurable current flows through the two dots.

ESR spectroscopy

The resonant ESR response is clearly observed in the transport measurements as a function of magnetic field (Fig. 2a, b), where satellite peaks develop at the resonant field $B_{ext} = \pm hf_{ac}/g\mu_B$ when the RF source is turned on (the zero-field peak arises from the inhomogeneous nuclear field, which admixes all the triplets with the singlet^{36,37}). The key signature of ESR is the linear dependence of the satellite peak location on the RF frequency, which is clearly seen in the data of Fig. 2c, where the RF frequency is varied from 10 to 750 MHz. From a linear fit through the top of the peaks we obtain a g -factor with modulus 0.35 ± 0.01 , which lies within the range of reported values for confined electron spins in GaAs quantum dots^{11,38–40}. We also verified explicitly that the resonance we observe is magnetic in origin and not caused by the electric field that the CPS generates as well; negligible response was observed when RF power is applied to the right side gate, generating mostly a RF electric field (see Supplementary Fig. S3).

The amplitude of the peaks in Fig. 2b increases linearly with RF power ($\sim B_{ac}^2$) before saturation occurs, as predicted²⁵ (Fig. 2b, inset). The ESR satellite peak is expected to be broadened by either the

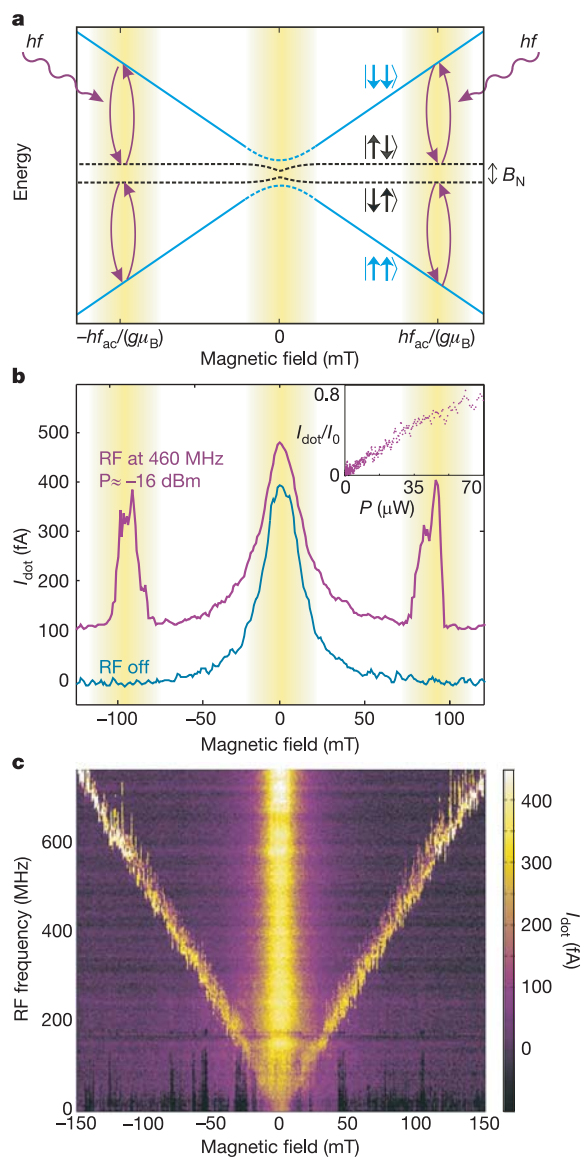


Figure 2 | ESR spin state spectroscopy. **a**, Energy diagram showing the relevant eigenstates of two electron spins in a double-dot, subject to an external magnetic field and nuclear fields. Because the nuclear field is generally inhomogeneous, the Zeeman energy is different in the two dots and results therefore in a different energy for $\uparrow\downarrow$ and $\downarrow\downarrow$, depending on the nuclear fields in the two dots. The yellow bands denote the ranges in B_{ext} where spin blockade is lifted (by the nuclear field or ESR) and current will flow through the dots. **b**, Current measured through the double-dot in the spin blockade regime, with (red trace, offset by 100 fA for clarity) and without (blue trace) a RF magnetic field. Satellite peaks appear as the external magnetic field is swept through the spin resonance condition. Each measurement point is averaged for one second, and is therefore expected to represent an average response over many nuclear configurations. The RF power P applied to the CPS is estimated from the power applied to the coax line and the attenuation in the lines. Inset, satellite peak height versus RF power ($f = 408$ MHz, $B_{\text{ext}} = 70$ mT, taken at slightly different gate voltage settings). The current is normalized to the current at $B_{\text{ext}} = 0$ ($= I_0$). Unwanted electric field effects are reduced by applying a compensating signal to the right side gate with opposite phase as the signal on the stripline (see Supplementary Fig. S4). This allowed us to obtain this curve up to relatively high RF powers. **c**, Current through the dots when sweeping the RF frequency and stepping the magnetic field. The ESR satellite peak is already visible at a small magnetic field of 20 mT and RF excitation of 100 MHz, and its location evolves linearly in field when increasing the frequency. For higher frequencies the satellite peak is broadened asymmetrically for certain sweeps, visible as vertical stripes. This broadening is time dependent, hysteretic in sweep direction, and changes with the dot level alignment. The horizontal line at 180 MHz is due to a resonance in the transmission line inside the dilution refrigerator.

excitation amplitude B_{ac} or incoherent processes, like cotunnelling, inelastic transitions (to the $S(0,2)$ state) or the statistical fluctuations in the nuclear field, whichever of the four has the largest contribution. No dependence of the width on RF power was found within the experimentally accessible range ($B_{\text{ac}} < 2$ mT). Furthermore, we suspect that the broadening is not dominated by cotunnelling or inelastic transitions because the corresponding rates are smaller than the observed broadening (see Supplementary Figs S4b and S2d). The observed ESR peaks are steeper on the flanks and broader than expected from the nuclear field fluctuations. In many cases, the peak width and position are even hysteretic in the sweep direction, suggesting that the resonance condition is shifted during the field sweep. We speculate that dynamic nuclear polarization due to feedback of the electron transport on the nuclear spins plays a central part here³⁷.

Coherent Rabi oscillations

Following the observation of magnetically induced spin flips, we next test whether we can also coherently rotate the spin by applying RF bursts with variable length. In contrast to the continuous-wave experiment, where detection and spin rotation occur at the same time, we pulse the system into Coulomb blockade during the spin manipulation. This eliminates decoherence induced by tunnel events from the left to the right dot during the spin rotations. The experiment consists of three stages (Fig. 3): initialization through spin blockade in a statistical mixture of $\uparrow\uparrow$ and $\downarrow\downarrow$, manipulation by a RF burst in Coulomb blockade, and detection by pulsing back for projection (onto $S(0,2)$) and tunnelling to the lead. When one of the electrons is rotated over $(2n + 1)\pi$ (with integer n), the two-electron state evolves to $\uparrow\downarrow$ (or $\downarrow\uparrow$), giving a maximum contribution to the current (as before, when the two spins are anti-parallel, one electron charge moves through the dots). However, no electron flow is expected after rotations of $2\pi n$, where one would find two parallel spins in the two dots after the RF burst.

We observe that the dot current oscillates periodically with the RF burst length (Fig. 4). This oscillation indicates that we performed driven, coherent electron spin rotations, or Rabi oscillations. A key characteristic of the Rabi process is a linear dependence of the Rabi frequency on the RF burst amplitude, B_{ac} ($f_{\text{Rabi}} = g\mu_B B_1/\hbar$ with $B_1 = B_{\text{ac}}/2$ due to the rotating wave approximation). We verify this by extracting the Rabi frequency from a fit of the current oscillations of Fig. 4b with a sinusoid, which gives the expected linear behaviour

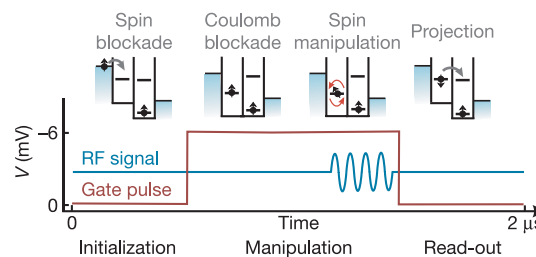


Figure 3 | The control cycle for coherent manipulation of the electron spin. During the ‘initialization’ stage the double-dot is tuned in the spin blockade regime. Electrons will move from left to right until the system is blocked with two parallel spins (either $\uparrow\uparrow$ or $\downarrow\downarrow$; in the figure only the $\uparrow\uparrow$ case is shown). For the ‘manipulation’ stage, the right dot potential is pulsed up so none of the levels in the right dot are accessible (Coulomb blockade), and a RF burst with a variable duration is applied. ‘Read-out’ of the spin state at the end of the manipulation stage is done by pulsing the right dot potential back; electron tunnelling to the right lead will then take place only if the spins were anti-parallel. The duration of the read-out and initialization stages combined was 1 μs , long enough ($1 \mu\text{s} > 1/T_L, 1/T_M, 1/T_R$) to have parallel spins in the dots at the end of the initialization stage with near certainty (this is checked by signal saturation when the pulse duration is prolonged). The duration of the manipulation stage is also held fixed at 1 μs to keep the number of pulses per second constant. The RF burst is applied just before the read-out stage starts.

(Fig. 4b, inset). From the fit we obtain $B_{ac} = 0.59$ mT for a stripline current I_{CPS} of ~ 1 mA, which agrees well with predictions from numerical finite element simulations (see Supplementary Fig. S1). The maximum B_1 we could reach in the experiment before electric field effects hindered the measurement was 1.9 mT, corresponding to $\pi/2$ rotations of only 27 ns (that is, a Rabi period of 108 ns, see Fig. 4b). If the accompanying electric fields from the stripline excitation could be reduced in future experiments (for example, by improving the impedance matching from coax to CPS), considerably faster Rabi flopping should be attainable.

The oscillations in Fig. 4b remain visible throughout the entire measurement range, up to 1 μ s. This is striking, because the Rabi period of ~ 100 ns is much longer than the time-averaged coherence time T_2^* of 10–20 ns (refs 14, 19, 35, 36) caused by the nuclear field fluctuations. The slow damping of the oscillations is only possible because the nuclear field fluctuates very slowly compared to the timescale of spin rotations and because other mechanisms, such as

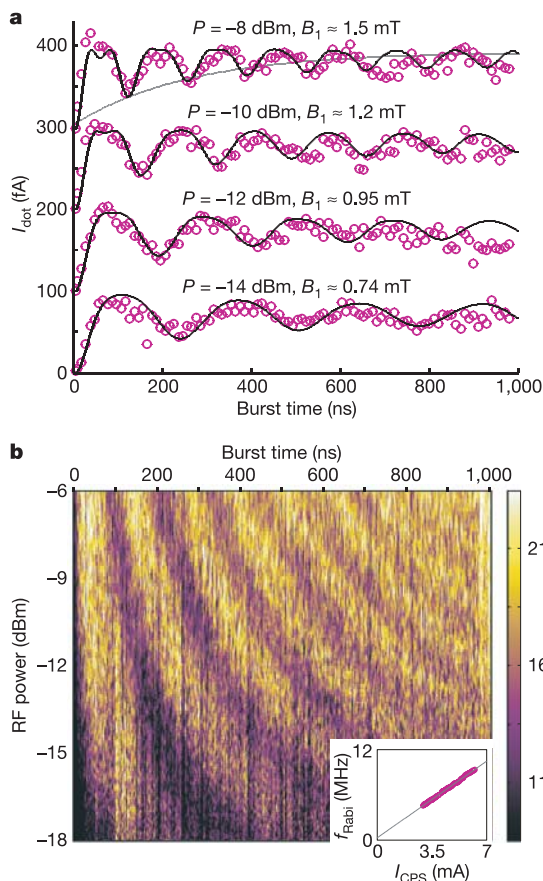


Figure 4 | Coherent spin rotations. **a**, The dot current—reflecting the spin state at the end of the RF burst—oscillates as a function of RF burst length (curves offset by 100 fA for clarity). The frequency of B_{ac} is set at the spin resonance frequency of 200 MHz ($B_{ext} = 41$ mT). The period of the oscillation increases and is more strongly damped for decreasing RF power. The RF power P applied to the CPS is estimated from the power applied to the coax line and the attenuation in the lines and RF switch. From P , the stripline current is calculated via the relation $P = \frac{1}{2} \left(\frac{I_{CPS}}{2} \right)^2 Z_0$ assuming perfect reflection of the RF wave at the short. Each measurement point is averaged over 15 s. We correct for a current offset which is measured with the RF frequency off-resonance (280 MHz). The solid lines are obtained from numerical computation of the time evolution, as discussed in the text. The grey line corresponds to an exponentially damped envelope. **b**, The oscillating dot current (represented in colourscale) is displayed over a wide range of RF powers (the sweep axis) and burst durations. The dependence of the Rabi frequency f_{Rabi} on RF power is shown in the inset. f_{Rabi} is extracted from a sinusoidal fit with the current oscillations from 10 to 500 ns for RF powers ranging from -12.5 dBm up to -6 dBm.

the spin-orbit interaction, disturb the electron spin coherence only on even longer timescales^{13,41,42}. We also note that the decay is not exponential (grey line in Fig. 4a), which is related to the fact that the nuclear bath is non-markovian (it has a long memory)⁴³.

Theoretical model

To understand better the amplitudes and decay times of the oscillations, we model the time evolution of the spins throughout the burst duration. The model uses a hamiltonian that includes the Zeeman splitting for the two spins and the RF field, which we take to be of equal amplitude in both dots (S_L and S_R refer to the electron spins in the left and right dot respectively):

$$H = g\mu_B(\mathbf{B}_{ext} + \mathbf{B}_{L,N})S_L + g\mu_B(\mathbf{B}_{ext} + \mathbf{B}_{R,N})S_R + g\mu_B \cos(\omega t)B_{ac}(S_L + S_R)$$

where $\mathbf{B}_{L,N}$ and $\mathbf{B}_{R,N}$ correspond to a single frozen configuration of the nuclear field in the left and right dot. This is justified because the electron spin dynamics is much faster than the dynamics of the nuclear system. From the resulting time evolution operator and assuming that the initial state is a statistical mixture of $\uparrow\uparrow$ and $\downarrow\downarrow$, we can numerically obtain the probability for having anti-parallel spins after the RF burst. This is also the probability that the left electron tunnels to the right dot during the read-out stage.

In the current measurements of Fig. 4a, each data point is averaged over 15 s, which presumably represents an average over many nuclear configurations. We include this averaging over different nuclear configurations in the model by taking 2,000 samples from a gaussian distribution of nuclear fields (with standard deviation $\sigma = \sqrt{\langle B_N^2 \rangle}$), and computing the probability that an electron tunnels out after the RF burst. When the electron tunnels, one or more additional electrons, say m , may subsequently tunnel through before $\uparrow\uparrow$ or $\downarrow\downarrow$ is formed and the current is blocked again. Taking m and σ as fitting parameters, we find good agreement with the data for $m=1.5$ and $\sigma = 2.2$ mT (solid black lines in Fig. 4a). This value for σ is comparable to that found in refs 35 and 36. The value found for m is different from what we would expect from a simple picture where all four spin states are formed with equal probability during the initialization stage, which would give $m = 1$. We do not understand this discrepancy, but it could be due to different tunnel rates for $\uparrow$ and $\downarrow$ or more subtle details in the transport cycle that we have neglected in the model.

Time evolution of the spin states during RF bursts

We now discuss in more detail the time evolution of the two spins during a RF burst. The resonance condition in each dot depends on the effective nuclear field, which needs to be added vectorially to B_{ext} . Through their continuous reorientation, the nuclear spins will bring the respective electron spins in the two dots on and off resonance as time progresses.

When a RF burst is applied to two spins initially in $\uparrow\uparrow$, and is on-resonance with the right spin only, the spins evolve as:

$$\begin{aligned} |\uparrow\rangle|\uparrow\rangle &\rightarrow |\uparrow\rangle\frac{|\uparrow\rangle+|\downarrow\rangle}{\sqrt{2}} \rightarrow |\uparrow\rangle|\downarrow\rangle \rightarrow \\ &|\uparrow\rangle\frac{|\uparrow\rangle-|\downarrow\rangle}{\sqrt{2}} \rightarrow |\uparrow\rangle|\uparrow\rangle \end{aligned}$$

When the RF burst is on-resonance with both spins, the time evolution is:

$$\begin{aligned} |\uparrow\rangle|\uparrow\rangle &\rightarrow \frac{|\uparrow\rangle+|\downarrow\rangle}{\sqrt{2}}\frac{|\uparrow\rangle+|\downarrow\rangle}{\sqrt{2}} \rightarrow |\downarrow\rangle|\downarrow\rangle \rightarrow \\ &\frac{|\uparrow\rangle-|\downarrow\rangle}{\sqrt{2}}\frac{|\uparrow\rangle-|\downarrow\rangle}{\sqrt{2}} \rightarrow |\uparrow\rangle|\uparrow\rangle \end{aligned}$$

In both cases, the RF causes transitions between the $\uparrow$ and $\downarrow$ states of single spin-half particles. When the RF is on-resonance with both spins, such single-spin rotations take place for both spins simultaneously. Because the current through the dots is proportional to the $S_z = 0$ probability ($\uparrow\downarrow$ or $\downarrow\uparrow$), we see that when both spins are excited simultaneously, the current through the dots will oscillate twice as fast as when only one spin is excited, but with only half the amplitude.

In the experiment, the excitation is on-resonance with only one spin at a time for most of the frozen nuclear configurations (Fig. 5). Only at the highest powers ($B_1/\sqrt{\langle B_{N,z}^2 \rangle} > 1$), both spins may be excited simultaneously (but independently) and a small double Rabi frequency contribution is expected, although it could not be observed, owing to the measurement noise.

Quantum gate fidelity

We can estimate the angle over which the electron spins are rotated in the Bloch sphere based on our knowledge of B_1 and the nuclear field fluctuations in the z -direction, again using the hamiltonian H . For the maximum ratio of $B_1/\sqrt{\langle B_{N,z}^2 \rangle} = B_1/(\sigma/\sqrt{3}) = 1.5$ reached in the present experiment, we achieve an average tip angle of 131° for an intended 180° rotation, corresponding to a fidelity of 73% (Fig. 5). Apart from using a stronger B_1 , the tip angle can be increased considerably by taking advantage of the long timescale of the nuclear field fluctuations. First, application of composite pulses, widely used in nuclear magnetic resonance to compensate for resonance off-sets⁴⁴, can greatly improve the quality of the rotations. A second solution comprises a measurement of the nuclear field (nuclear state narrowing^{45–47}), so that the uncertainty in the nuclear field is reduced, and accurate rotations can be realized for as long as the nuclear field remains constant.

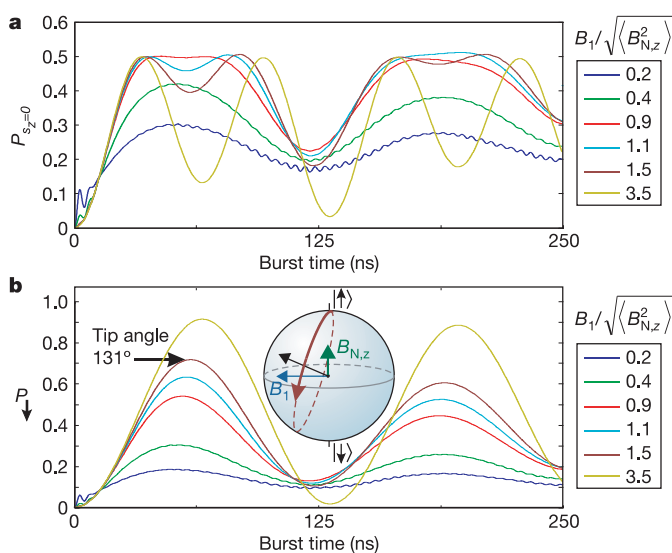


Figure 5 | Time evolution of the spin states. **a**, Probability for the two spins to be in $\uparrow\downarrow$ or $\downarrow\uparrow$ ($S_z = 0$) at the end of a RF burst, with initial state $\uparrow\uparrow$, computed using the hamiltonian H presented in the main text, for six different values of $\sigma_{N,z} = \langle B_{N,z}^2 \rangle^{1/2}$ (fixed $B_1 = 1.5$ mT, $B_{\text{ext}} = 40$ mT, each of the traces is averaged over 2,000 static nuclear configurations). As expected, the oscillation contains a single frequency for B_1 small compared to $\sigma_{N,z}$, corresponding to the Rabi oscillation of a single spin. The oscillation develops a second frequency component, twice as fast as the first, when $B_1/\sigma_{N,z} > 1$. For $B_1/\sigma_{N,z} > 4$ the double frequency component is dominant, reflecting the simultaneous Rabi oscillation of the two spins. **b**, Probability for one of the spins to be $\downarrow$ at the end of a RF burst. The spin state evolution is computed as in **a**. This oscillation represents the Rabi oscillation of one spin by itself. For increasing B_1 , the maximum angle over which the spin is rotated in the Bloch sphere increases as well. In the experiment, this angle could not be measured directly, because the current measurement constitutes a two-spin measurement, not a single-spin measurement. We can, however, extract the tip angle from **P**.

In future experiments, controllable addressing of the spins in the two dots separately can be achieved through a gradient in either the static or the oscillating magnetic field. Such gradient fields can be created relatively easily using a ferromagnet or an asymmetric stripline. Alternatively, the resonance frequency of the spins can be selectively shifted using local g -factor engineering^{48,49}. The single spin rotations reported here, in combination with single-shot spin read-out^{13,18} and the tunable exchange coupling in double dots¹⁴, offers many new opportunities, such as measuring the violation of Bell's inequalities or the implementation of simple quantum algorithms.

Received 26 April; accepted 6 July 2006.

- Nielsen, M. A. & Chuang, I. L. *Quantum Computation and Quantum Information* (Cambridge Univ. Press, Cambridge, 2000).
- Shor, P. W. in *Proc. 35th Annu. Symp. on the Foundations of Computer Science* (ed. Goldwasser, S.) 124–134 (IEEE Computer Society Press, Los Alamitos, California, 1994).
- Zoller, P. et al. Quantum information processing and communication, Strategic report on current status, visions and goals for research in Europe. *Eur. Phys. J. D* **36**, 203–228 (2005).
- DiVincenzo, D. P. Quantum computation. *Science* **270**, 255–261 (1995).
- Loss, D. & DiVincenzo, D. P. Quantum computation with quantum dots. *Phys. Rev. A* **57**, 120–126 (1998).
- Austing, D. G., Honda, T., Muraki, K., Tokura, Y. & Tarucha, S. Quantum dot molecules. *Phys. B Cond. Matter* **249–251**, 206–209 (1998).
- Ciorga, M. et al. Addition spectrum of a lateral dot from Coulomb and spin-blockade spectroscopy. *Phys. Rev. B* **61**, R16315 (2000).
- Elzerman, J. M. et al. Few-electron quantum dot circuit with integrated charge read out. *Phys. Rev. B* **67**, 161308 (2003).
- Bayer, M. et al. Coupling and entangling of quantum states in quantum dot molecules. *Science* **291**, 451–453 (2001).
- Atature, M. et al. Quantum-dot spin-state preparation with near-unity fidelity. *Science* **312**, 551–553 (2006).
- Hanson, R. et al. Zeeman energy and spin relaxation in a one-electron quantum dot. *Phys. Rev. Lett.* **91**, 196802 (2003).
- Fujisawa, T., Austing, D. G., Tokura, Y., Hirayama, Y. & Tarucha, S. Allowed and forbidden transitions in artificial hydrogen and helium atoms. *Nature* **419**, 278–281 (2002).
- Elzerman, J. M. et al. Single-shot read-out of an individual electron spin in a quantum dot. *Nature* **430**, 431–435 (2004).
- Petta, J. R. et al. Coherent manipulation of coupled electron spins in semiconductor quantum dots. *Science* **309**, 2180–2184 (2005).
- Schleser, R. et al. Time-resolved detection of individual electrons in a quantum dot. *Appl. Phys. Lett.* **85**, 2005–2007 (2004).
- Vandersypen, L. M. K. et al. Real-time detection of single-electron tunneling using a quantum point contact. *Appl. Phys. Lett.* **85**, 4394–4396 (2004).
- Lu, W., Ji, Z. Q., Pfeiffer, L., West, K. W. & Rimbey, A. J. Real-time detection of electron tunnelling in a quantum dot. *Nature* **423**, 422–425 (2003).
- Hanson, R. et al. Single-shot readout of electron spin states in a quantum dot using spin-dependent tunnel rates. *Phys. Rev. Lett.* **94**, 196802 (2005).
- Dutt, M. V. G. et al. Stimulated and spontaneous optical generation of electron spin coherence in charged GaAs quantum dots. *Phys. Rev. Lett.* **94**, 227403 (2005).
- Grelich, A. et al. Optical control of spin coherence in singly charged (In,Ga)As/GaAs quantum dots. *Phys. Rev. Lett.* **96**, 227401 (2006).
- Poole, C. P. *Electron Spin Resonance* 2nd edn (Wiley, New York, 1983).
- Xiao, M., Martin, I., Yablonovitch, E. & Jiang, H. W. Electrical detection of the spin resonance of a single electron in a silicon field-effect transistor. *Nature* **430**, 435–439 (2004).
- Jelesko, F., Gaebel, T., Popa, I., Gruber, A. & Wrachtrup, J. Observation of coherent oscillations in a single electron spin. *Phys. Rev. Lett.* **92**, 076401 (2004).
- Rugar, D., Budakian, R., Mamin, H. J. & Chui, B. W. Single spin detection by magnetic resonance force microscopy. *Nature* **430**, 329–332 (2004).
- Engel, H. A. & Loss, D. Detection of single spin decoherence in a quantum dot via charge currents. *Phys. Rev. Lett.* **86**, 4648–4651 (2001).
- Imamoglu, A. et al. Quantum information processing using quantum dot spins and cavity QED. *Phys. Rev. Lett.* **83**, 4204–4207 (1999).
- Kato, Y., Myers, R. C., Gossard, A. C. & Awschalom, D. D. Coherent spin manipulation without magnetic fields in strained semiconductors. *Nature* **427**, 50–53 (2003).
- Golovach, V. N., Borhani, M. & Loss, D. Electric dipole induced spin resonance in quantum dots. Preprint at (www.arXiv.org/cond-mat/0601674) (2006).
- Tokura, Y., Van der Wiel, W. G., Obata, T. & Tarucha, S. Coherent single electron spin control in a slanting Zeeman field. *Phys. Rev. Lett.* **96**, 047202 (2006).
- Ono, K., Austing, D. G., Tokura, Y. & Tarucha, S. Current rectification by Pauli exclusion in a weakly coupled double quantum dot system. *Science* **297**, 1313–1317 (2002).

31. Engel, H. A. *et al.* Measurement efficiency and n-shot readout of spin qubits. *Phys. Rev. Lett.* **93**, 106804 (2004).
32. Blaauboer, M. & DiVincenzo, D. P. Detecting entanglement using a double-quantum-dot turnstile. *Phys. Rev. Lett.* **95**, 160402 (2005).
33. Engel, H. A. & Loss, D. Fermionic bell-state analyzer for spin qubits. *Science* **309**, 586–588 (2005).
34. Johnson, A. C., Petta, J. R., Marcus, C. M., Hanson, M. P. & Gossard, A. C. Singlet-triplet spin blockade and charge sensing in a few-electron double quantum dot. *Phys. Rev. B* **72**, 165308 (2005).
35. Johnson, A. C. *et al.* Triplet-singlet spin relaxation via nuclei in a double quantum dot. *Nature* **435**, 925–928 (2005).
36. Koppens, F. H. L. *et al.* Control and detection of singlet-triplet mixing in a random nuclear field. *Science* **309**, 1346–1350 (2005).
37. Jouravlev, O. N. & Nazarov, Y. V. Electron transport in a double quantum dot governed by a nuclear magnetic field. *Phys. Rev. Lett.* **96**, 176804 (2006).
38. Potok, R. M. *et al.* Spin and polarized current from Coulomb blockaded quantum dots. *Phys. Rev. Lett.* **91**, 016802 (2003).
39. Willems van Beveren, L. H. W. *et al.* Spin filling of a quantum dot derived from excited-state spectroscopy. *New J. Phys.* **7**, 182 (2005).
40. Kogan, A. *et al.* Measurements of Kondo and spin splitting in single-electron transistors. *Phys. Rev. Lett.* **93**, 166602 (2004).
41. Kroutvar, M. *et al.* Optically programmable electron spin memory using semiconductor quantum dots. *Nature* **432**, 81–84 (2004).
42. Golovach, V. N., Khaetskii, A. & Loss, D. Phonon-induced decay of the electron spin in quantum dots. *Phys. Rev. Lett.* **93**, 016601 (2004).
43. Coish, W. A. & Loss, D. Hyperfine interaction in a quantum dot: Non-Markovian electron spin dynamics. *Phys. Rev. B* **70**, 195340 (2004).
44. Vandersypen, L. M. K. & Chuang, I. L. NMR techniques for quantum control and computation. *Rev. Mod. Phys.* **76**, 1037–1069 (2004).
45. Klauser, D., Coish, W. A. & Loss, D. Nuclear spin state narrowing via gate-controlled Rabi oscillations in a double quantum dot. *Phys. Rev. Lett.* **96**, 176804 (2006).
46. Giedke, G., Taylor, J. M., D'Alessandro, D., Lukin, D. & Imamoglu, A. Quantum measurement of the nuclear spin polarization in quantum dots. Preprint at www.arXiv.org/quant-ph/0508144 (2005).
47. Stepanenko, D., Burkard, G., Giedke, G. & Imamoglu, A. Enhancement of electron spin coherence by optical preparation of nuclear spins. *Phys. Rev. Lett.* **96**, 136401 (2006).
48. Salis, G. *et al.* Electrical control of spin coherence in semiconductor nanostructures. *Nature* **414**, 619–622 (2001).
49. Jiang, H. W. & Yablonovitch, E. Gate-controlled electron spin resonance in GaAs/Al_xGa_{1-x}As heterostructures. *Phys. Rev. B* **64**, 041307 (2001).
50. Holleitner, A. W., Blick, R. H. & Eberl, K. Fabrication of coupled quantum dots for multiport access. *Appl. Phys. Lett.* **82**, 1887–1889 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank W. Coish, J. Elzerman, D. Klauser, A. Lupascu, D. Loss and in particular J. Folk for discussions; R. Schouten, B. van der Enden and W. den Braver for technical assistance; The International Research Centre for Telecommunication and Radar at the Delft University of Technology for assistance with the stripline simulations. Supported by the Dutch Organization for Fundamental Research on Matter (FOM), the Netherlands Organization for Scientific Research (NWO) and the Defense Advanced Research Projects Agency Quantum Information Science and Technology programme.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to L.M.K.V. (l.m.k.vandersypen@tudelft.nl) and F.H.L.K. (f.h.l.koppens@tudelft.nl).

ARTICLES

A genomic code for nucleosome positioning

Eran Segal¹, Yvonne Fondufe-Mittendorf², Lingyi Chen², AnnChristine Thåström², Yair Field¹, Irene K. Moore², Ji-Ping Z. Wang³ & Jonathan Widom²

Eukaryotic genomes are packaged into nucleosome particles that occlude the DNA from interacting with most DNA binding proteins. Nucleosomes have higher affinity for particular DNA sequences, reflecting the ability of the sequence to bend sharply, as required by the nucleosome structure. However, it is not known whether these sequence preferences have a significant influence on nucleosome position *in vivo*, and thus regulate the access of other proteins to DNA. Here we isolated nucleosome-bound sequences at high resolution from yeast and used these sequences in a new computational approach to construct and validate experimentally a nucleosome–DNA interaction model, and to predict the genome-wide organization of nucleosomes. Our results demonstrate that genomes encode an intrinsic nucleosome organization and that this intrinsic organization can explain ~50% of the *in vivo* nucleosome positions. This nucleosome positioning code may facilitate specific chromosome functions including transcription factor binding, transcription initiation, and even remodelling of the nucleosomes themselves.

Eukaryotic genomic DNA exists as highly compacted nucleosome arrays called chromatin. Each nucleosome contains a 147-base-pair (bp) stretch of DNA, which is sharply bent and tightly wrapped around a histone protein octamer¹. This sharp bending occurs at every DNA helical repeat (~10 bp), when the major groove of the DNA faces inwards towards the histone octamer, and again ~5 bp away, with opposite direction, when the major groove faces outward. Bends of each direction are facilitated by specific dinucleotides^{2,3}. Neighbouring nucleosomes are separated from each other by 10–50-bp-long stretches of unwrapped linker DNA⁴; thus, 75–90% of genomic DNA is wrapped in nucleosomes. Access to DNA wrapped in a nucleosome is occluded¹ for polymerase, regulatory, repair and recombination complexes, yet nucleosomes also recruit other proteins through interactions with their histone tail domains⁵. Thus, the detailed locations of nucleosomes along the DNA may have important inhibitory or facilitatory roles^{6,7} in regulating gene expression.

DNA sequences differ greatly in their ability to bend sharply^{2,3,8}. Consequently, the ability of the histone octamer to wrap differing DNA sequences into nucleosomes is highly dependent on the specific DNA sequence^{9,10}. *In vitro* studies show this range of affinities to be 1,000-fold or greater¹¹. Thus, nucleosomes have substantial DNA sequence preferences. A key question is whether genomes use these sequence preferences to control the distribution of nucleosomes *in vivo* in a way that strongly impacts on the ability of DNA binding proteins to access particular binding sites. By controlling binding site accessibility in this way, genomes could, for example, target the binding of transcription factors towards appropriate sites and away from irrelevant, non-functional sites⁹.

One view is that the sequence preferences of nucleosomes might not be meaningful. Nucleosome positions might be regulated in cells *in trans* by the abundant¹² ATP-dependent nucleosome remodelling complexes¹³, which might over-ride the sequence preferences of nucleosomes and move them to new locations whenever needed. Another view, however, is that remodelling factors do not themselves

determine the destinations of the nucleosomes that they mobilize. Rather, the remodelling complexes may allow nucleosomes to sample alternative positions rapidly, resulting in a thermodynamic equilibrium between the nucleosomes and the site-specific DNA binding proteins that compete with nucleosomes for occupancy along the genome. In this view, nucleosome positions are regulated *in cis* by their intrinsic sequence preferences, which would then have significant regulatory roles. In this *cis* regulation model, we expect the genome to encode a nucleosome organization, intrinsic to the DNA sequence alone, comprising sequences with both low and high affinity for nucleosomes. Many of the high-affinity sequences should then be occupied by nucleosomes *in vivo*. Moreover, the detailed distribution of nucleosome positions encoded by the genome should significantly influence chromosome functions genome-wide.

Here we report the results of a combined experimental and computational approach to detect the DNA sequence preferences of nucleosomes and the intrinsic nucleosome organization of the genome that these preferences dictate. Our findings demonstrate that eukaryotic genomes use a nucleosome positioning code, and link the resulting nucleosome positions to specific chromosome functions.

Validating a nucleosome–DNA interaction model

To construct a model for nucleosome–DNA interactions in yeast (Fig. 1a), we used a genome-wide assay to isolate DNA regions that were stably wrapped in nucleosomes. Our experimental method maps nucleosomes on the yeast genome with greater accuracy than previous approaches, resulting in a set of 199 mononucleosome DNA sequences of length 142–152 bp (Supplementary Fig. 1). We used this collection of sequences to construct a probabilistic model that represents the DNA sequence preferences of yeast nucleosomes (Supplementary Fig. 2). Our approach resembles that used for representing the binding specificities of transcription factors from a collection of known sites, but with two main distinctions: first, in contrast to the mononucleotide probability distributions used for

¹Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 76100, Israel. ²Department of Biochemistry, Molecular Biology and Cell Biology, Northwestern University, 2153 Sheridan Road, Evanston, Illinois 60208, USA. ³Department of Statistics, Northwestern University, 2006 Sheridan Road, Evanston, Illinois 60208, USA.

transcription factors, we use dinucleotide probability distributions (as dinucleotides are the simplest sequence elements to capture the sequence-dependent mechanics of DNA bending¹⁴ that are essential for histone–DNA association³); second, when constructing the model we represent the two-fold symmetry axis of the nucleosome structure¹ by including the reverse complement of each sequence in the nucleosome collection. More sophisticated nucleosome–DNA interaction models based on mixture models¹⁵ or on the expectation-maximization algorithm¹⁶ yielded equivalent results.

As expected for a nucleosome–DNA interaction model, the resulting model exhibits distinctive sequence motifs that recur periodically at the DNA helical repeat and are known to facilitate the sharp bending of DNA around the nucleosome³. These include ~10-bp periodic AA/TT/TA dinucleotides that oscillate in phase with each other (Fig. 1b) and out of phase with ~10-bp periodic GC dinucleotides. Moreover, the same periodicities and phase relationships were derived independently from a collection of 177 natural nucleosomes from chicken², and they arose again in three independent *in vitro* experiments that selected for stable nucleosomes. These *in vitro* selection experiments include one on chemically synthesized random DNA¹⁷, one on mouse genomic DNA¹⁸, and a new experiment that we performed on yeast genomic DNA (see Methods). The similarities among these independently derived nucleosome patterns are striking and quantitatively significant (Fig. 1b and Supplementary Figs 3–5), for example, $P < 10^{-50}$ for yeast–chicken *in vivo* similarity.

We experimentally validated the importance of these periodic sequence motifs for nucleosome–DNA interactions *in vitro*. Improving

the agreement of a sequence with these motifs increased its binding affinity to the nucleosome, whereas changing the periodicity or deleting the key motifs decreased that affinity (Fig. 1c–e and Supplementary Fig. 6). In addition, these periodic motifs did not arise in alignments of randomly chosen regions in the yeast or chicken genomes (Supplementary Fig. 7). Together, these results establish that the distinctive motifs in our model represent DNA sequence preferences of nucleosomes (Fig. 1f).

If genomes use these sequence preferences, then high-affinity sequences should be prevalent in the genome. Indeed, we found that intergenic and coding regions in the yeast genome contain many more high-affinity DNA sequences than expected by chance ($P < 10^{-200}$ for both intergenic and coding regions; Supplementary Fig. 8), and that scores at positions separated by 10 bp are strongly correlated (Supplementary Fig. 9). Together with the distinctive features of the yeast *in vivo* nucleosome collection, these results show that sequence motifs for positioning nucleosomes are abundantly encoded in the yeast genome and that nucleosomes occupy these sequences *in vivo*.

Predicting nucleosome organization in genomic DNA sequence

We next sought to understand how the encoded nucleosome preferences integrate to specify the intrinsic genome-wide positioning of nucleosomes. This task is non-trivial because encoded nucleosome positions are correlated through steric hindrance. We designed a thermodynamic model that defines an apparent free energy for every organization of nucleosomes on the DNA, taking steric hindrance and competition between nucleosomes into account (see Methods).

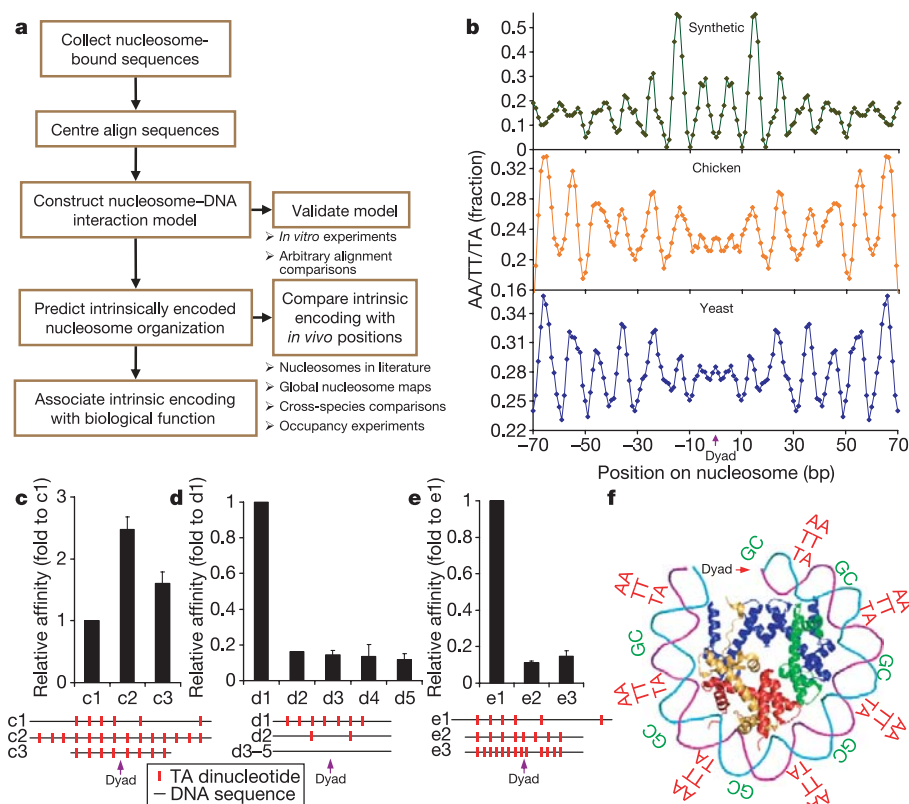


Figure 1 | Probabilistic nucleosome–DNA interaction model. **a**, Flow chart illustrating our approach. **b**, Fraction (3-bp moving average) of AA/TT/TA dinucleotides at each position of centre-aligned yeast, chicken² or random synthetic sequences, showing ~10-bp periodicity of these dinucleotides. **c–e**, *In vitro* experiments. Positions of the key AA/TT/TA dinucleotides on the tested sequences are indicated. Error bars are s.e.m. **c**, Nucleosome binding affinities of sequences

c2 and c3 (ref. 44), which include additional dinucleotide motifs at key positions, relative to the affinity of c1. **d**, Sequences d2–d5 have dinucleotide motifs removed from key positions in e1. **e**, Sequences e2 and e3 have disrupted spacing between the key dinucleotide motifs. **f**, Key dinucleotides inferred from the alignments are shown relative to the three-dimensional structure of one-half of the symmetric nucleosome.

A dynamic programming method¹⁹ evaluated efficiently all sterically allowed organizations, yielding both the probability that each base pair is occupied by any nucleosome (average nucleosome occupancy) and the genomic locations of the sites at which nucleosomes have a high probability of starting (stably positioned nucleosomes).

The resulting intrinsic nucleosome organization differs qualitatively at different genomic locations. In some cases, several mutually exclusive organizations dominate (Supplementary Fig. 10a, b); in others, a single organization dominates (Supplementary Fig. 10c); and yet in others no particular organization dominates (Supplementary Fig. 10d). Comparing these diverse intrinsic organizations to known transcription factor binding sites²⁰ reveals the potential regulatory role of nucleosomes: nucleosomes may have a strong affinity to occupy transcription factor binding sites (rendering them inaccessible) in some genomic locations (Supplementary Fig. 10a), but a weak affinity to occupy sites (thereby increasing their accessibility) in other locations (Supplementary Fig. 10b).

Predicted nucleosome organization reflects *in vivo* data

By comparing actual *in vivo* nucleosome positions to our predicted or experimentally measured intrinsically encoded positions, we can test whether *in vivo* positions are dictated by the genomic sequence. To this end, we used five different approaches. First, we measured the distance between our predicted stable nucleosome positions (stability probability ≥ 0.2 ; see Methods) and 99 experimentally mapped nucleosome positions at 11 loci^{21–28} (Supplementary Fig. 11). There is some disagreement between different experimental measurements of nucleosome positions (Fig. 2b and Supplementary Fig. 12), hence discrepancies between our predictions and literature reports are attributable to inaccuracies both in our model and in the literature. Even so, six loci showed substantial correspondence (Fig. 2 and Supplementary Figs 13–22). Overall, 54% of our predicted stable nucleosomes were within 35 bp of the literature positions, significantly more than the $39 \pm 1\%$ expected by chance ($P < 10^{-16}$).

Second, we compared our predictions to three genome-wide measurements of nucleosome positions at low^{29,30} or higher³¹ resolution. Our model showed significant correspondence to these experiments, predicting lower occupancy at nucleosome-depleted (low

nucleosome abundance) coding or intergenic regions^{29,30} (Supplementary Figs 23–25; 68% of 57 depleted coding regions and 76% of 294 depleted intergenic regions had predicted low occupancy compared with 30% ($P < 10^{-6}$) and 56% ($P < 10^{-9}$), respectively, expected by chance). The model also showed strong correspondence with the higher resolution nucleosome map³¹: 45% of our predicted stable nucleosomes were within 35 bp of experimentally determined nucleosome positions³¹ compared with $32 \pm 1\%$ expected by chance, $P < 10^{-15}$ (Supplementary Figs 26 and 27). Notably, our predictions also match closely the stereotyped chromatin organization at Pol II promoters as revealed by the higher resolution nucleosome map³¹, and the most stable nucleosome predicted by our model at promoters is located precisely (within 8 bp) where stable nucleosomes containing the histone variant H2A.Z are located *in vivo*³² (Fig. 5a).

Third, we compared the yeast model predictions to those of a model constructed independently using only nucleosome-bound sequences from chicken. The predictions of the chicken model when applied to the yeast genome correlated strongly with those of the yeast model (Supplementary Fig. 28) and with the genome-wide experimental measurements of nucleosome occupancy at yeast coding and intergenic regions^{29–31}: 35% of 57 depleted coding regions and 72% of 294 depleted intergenic regions had predicted low occupancy compared with 4% ($P < 10^{-4}$) and 53% ($P < 10^{-8}$) expected by chance.

Fourth, we carried out a new selection for nucleosome formation on yeast genomic DNA *in vitro*. This experiment directly reveals intrinsically encoded, individual high-affinity nucleosome positions. These *in vitro* nucleosome locations overlap significantly with our *in vivo* yeast nucleosome collection: 32% of 339 selected *in vitro* nucleosomes overlapping the *in vivo* bound sequences compared with 5% ($P < 10^{-5}$) expected by chance. The *in vitro* selected nucleosomes are particularly enriched in intergenic regions that have a high predicted nucleosome occupancy, compared with random genomic locations and to locations immediately upstream or downstream of the selected nucleosomes ($P < 10^{-3}$; Fig. 3c and Supplementary Figs 29 and 30).

Finally, we experimentally tested whether our highest occupancy predictions are highly occupied by nucleosomes *in vivo*, by measuring

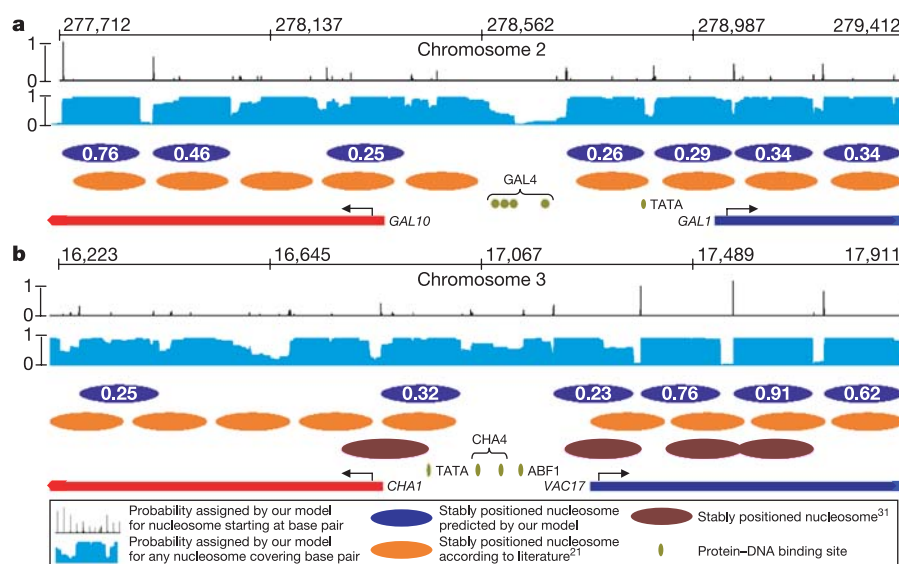


Figure 2 | Genome-wide prediction of intrinsic nucleosome organization and comparison to literature-reported, experimentally identified nucleosome positions. **a**, Detailed view of the *GAL1-10* locus, with literature-reported nucleosome positions²¹ (orange ovals). Black trace, probability of a nucleosome starting at each base pair; blue ovals, high probability nucleosomes predicted from our model (probability is

indicated); light-blue trace, average occupancy by any nucleosome at each base pair; red and blue bars, protein-coding regions; green ovals, conserved and bound DNA-binding sites²⁰. **b**, Same as in **a**, but for the *CHA1* locus²⁴; brown ovals, nucleosomes reported from other experiments³¹. The discrepancies between the two sets of literature-reported nucleosome positions highlight the uncertainty in such measurements.

their *in vivo* nucleosome occupancies and comparing them to the occupancies at three nucleosome sites flanking the *GAL1-10* and *PHO5* promoters for which the nucleosome positions are known. Five of the eight predictions tested yielded *in vivo* occupancies comparable to or greater than those of the known nucleosome positions (Fig. 3a), indicating that ~60% of the intrinsically high-occupancy nucleosome sites on the DNA sequence are strongly occupied *in vivo*. In 10 out of 11 cases, these predicted nucleosome positions also had higher occupancy than regions 73 bp (one-half the length of a nucleosome) upstream or downstream from the predicted position (Fig. 3b and Supplementary Figs 31 and 32).

Taken together, these results show that ~50% of the *in vivo* nucleosome organization can be explained solely by the sequence preferences of nucleosomes. Moreover, these results indicate that the nucleosome depletions observed at coding and intergenic regions^{29–31} are attributable in part to unstable nucleosomes (that is, positions on the DNA sequence that nucleosomes have a low probability of occupying) encoded in these regions.

Global features of intrinsic nucleosome organization in yeast

We next studied global properties of the intrinsic nucleosome organization in yeast. First, we examined the predicted stability of all 11,802,267 possible genome-wide nucleosome positions; 15,777 were highly stable (stability probability ≥ 0.5), significantly more than the $10,940 \pm 339$ ($P < 10^{-20}$) expected by chance. This result may indicate the existence of many genomic locations that encode highly stable nucleosomes, together covering 20% of the genome.

Second, we asked whether individual nucleosomes are organized into higher-ordered nucleosome arrays. The distribution of pairwise distances between positions of the highly stable nucleosomes revealed significant correlations persisting over at least six adjacent nucleosomes, with an average nucleosome repeat length of 177 bp

(Fig. 3d). We found similar strong correlations when considering the average nucleosome occupancy predictions (Supplementary Fig. 33). We conclude that the yeast genome not only encodes the preferred positions of individual nucleosomes, but also directly encodes higher structural levels of chromatin organization.

Nucleosome organization varies by type of genomic region

We next asked whether the genome's intrinsic encoding of nucleosome occupancy varies across different types of chromosomal regions, including centromeres, telomeres, intergenic and coding regions, and specific gene classes (Fig. 4a and Supplementary Fig. 34). Indeed, several types of regions had markedly high or low predicted occupancy. The highest predicted occupancy was over centromeres, indicating that centromere function requires enhanced stability of histone–DNA interactions that are encoded in the genomic sequence.

One might think that genomes would facilitate high gene expression levels by encoding unstable nucleosomes over highly expressed genes. Consistent with this expectation, the highly expressed ribosomal RNA and transfer RNA genes stood out as having markedly low predicted nucleosome occupancy.

In contrast to the ubiquitously expressed tRNAs, many other genes vary their expression between high and low levels in different conditions. However, as the genome sequence is static, it cannot simultaneously encode a nucleosome organization that would facilitate both high and low expression levels. Ribosomal proteins are one such example. Our model predicts high nucleosome occupancy encoded over these genes. Thus, the genome sequence does not facilitate the nucleosome depletion²⁹ and high expression of ribosomal proteins observed during normal growth, which therefore must be governed by other factors. Instead, the genome facilitates the rapid nucleosome reassembly²⁹ and strong repression of these genes observed under stress^{33,34}. These results show how the genome's

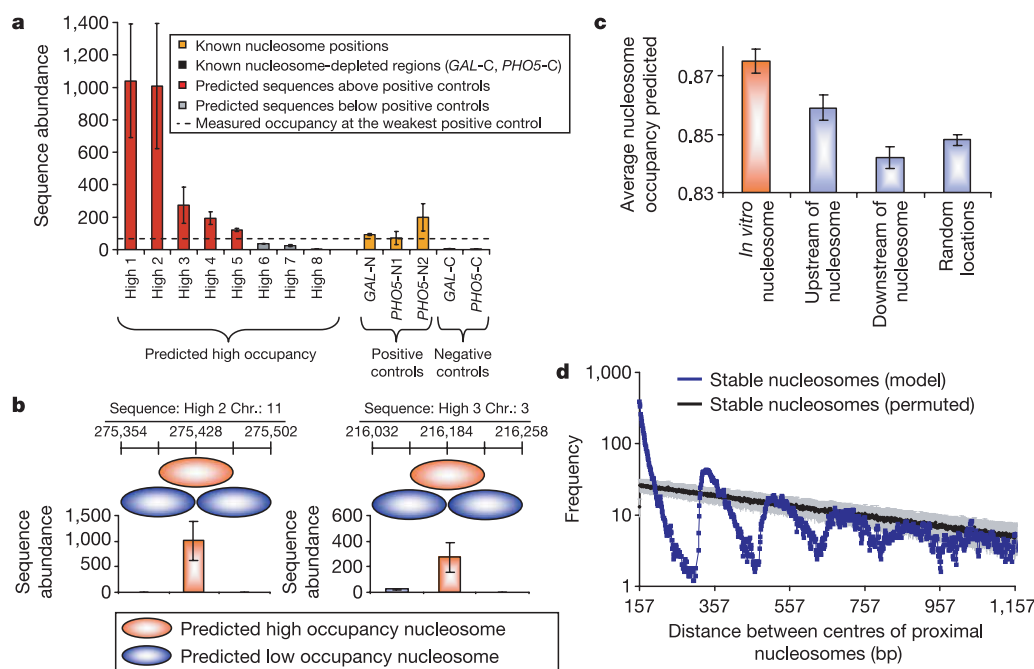


Figure 3 | Higher-order features of intrinsic nucleosome organization and comparison with *in vivo* occupancy experiments. **a**, Experimentally measured nucleosome occupancy *in vivo* for eight high-occupancy predictions, compared with high- and low-occupied locations in the *GAL1-10* and *PHO5* promoters. Error bars are s.d. **b**, *In vivo* nucleosome occupancy measured at predicted low-occupancy regions that are one-half nucleosome distance upstream and downstream (light blue) from the high-occupancy (orange) predictions of **a**. See Supplementary Fig. 31 for additional measurements. Results of **a** and **b** were consistent when

normalized for the sequence specificity of micrococcalnuclease (Supplementary Fig. 32). Error bars are s.d. **c**, Predicted nucleosome occupancy in intergenic regions for nucleosomes obtained from an *in vitro* selection experiment (orange) compared with predicted nucleosome occupancy in immediately upstream or downstream locations, or to random genomic locations (light blue). Error bars are s.e.m. **d**, Number of all pairs of proximal stable nucleosomes per centre-to-centre nucleosome distance, compared to the mean (black) and standard deviation (grey) in 100 permutations. Blue, yeast model (stability probability ≥ 0.5).

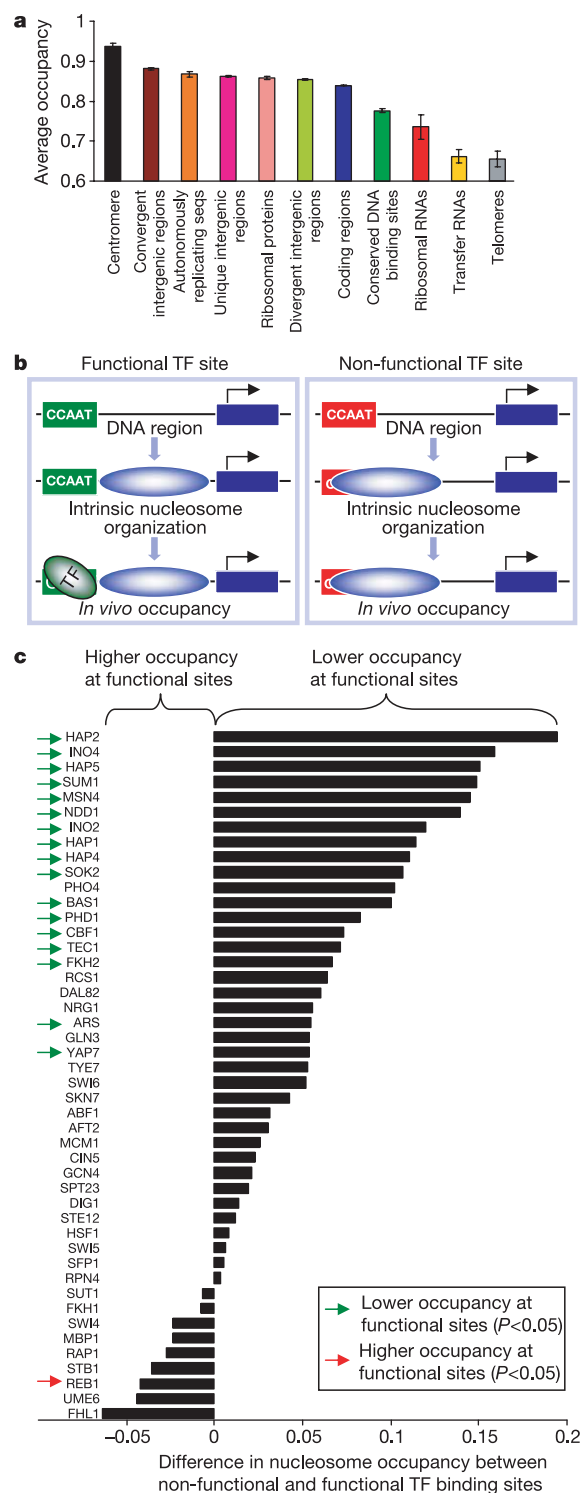


Figure 4 | Intrinsic nucleosome occupancy varies with genomic location type and is low at functional transcription factor binding sites. **a**, Average occupancies and standard errors for different types of genomic regions. **b**, Schematic illustrating how the intrinsic nucleosome organization may facilitate binding of transcription factors (TF) at functional sites, while disfavoring binding at identical non-functional sites that occur by chance. **c**, Difference in predicted nucleosome occupancy between non-functional and functional transcription factor binding sites (absolute occupancy levels are shown in Supplementary Fig. 36). Green arrows, 17 factors having significantly lower nucleosome occupancy at functional sites compared with non-functional sites²⁰; red arrow, 1 factor having significantly higher nucleosome occupancy at non-functional sites compared with functional sites.

statically encoded nucleosome organization may contribute to the dynamic process of gene regulation.

Nucleosomes facilitate their own remodelling

We tested whether the variation of nucleosome occupancy that we observed at different types of chromosomal region also extended to

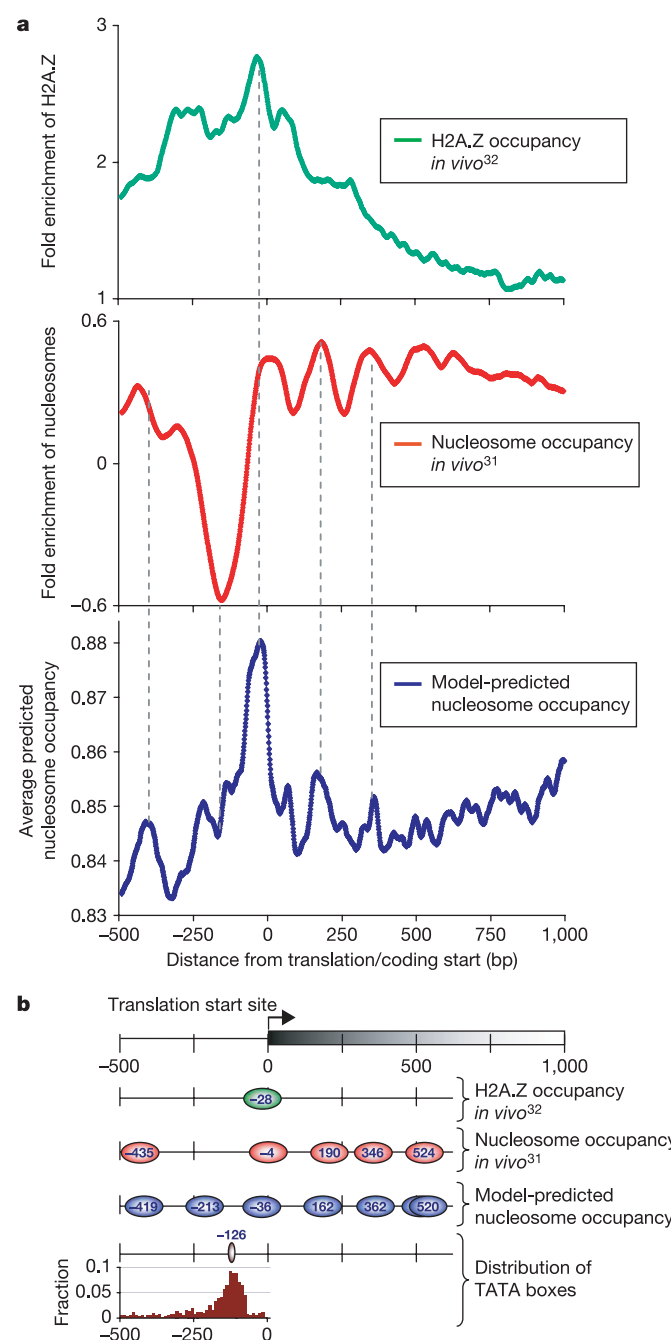


Figure 5 | Genomes encode unstable nucleosomes at transcriptional start sites. **a**, Average across all yeast genes of the *in vivo* occupancy of nucleosomes containing the histone variant H2A.Z³² (green) or of canonical nucleosomes³¹ (red), compared to the nucleosome occupancy predicted by our yeast nucleosome model (blue); all versus distance from the translation open reading frame (ORF) start site. The ORF-proximal peak of our model is statistically significant (Supplementary Fig. 37). **b**, The most probable nucleosome organization, based on **a**. Each nucleosome (ovals; labels represent nucleosome centres) is centred to a corresponding peak in **a**. Bottom graph shows distribution of TATA boxes⁴² relative to ORF start sites; brown oval is median TATA box location.

other sets of functionally related genes. We collected 1,949 different sets of yeast genes from a functional gene annotation database³⁵ and from a wide range of genomic studies^{20,36–40}, and found that indeed many gene sets showed a significant association with either high or low predicted nucleosome occupancy (Supplementary Fig. 35). Notably, of all gene sets tested, the most significant association predicted low occupancy at regions bound by the chromatin remodelling complex RSC⁴⁰ ($P < 10^{-34}$). This implies that genomes facilitate their own chromatin remodelling by encoding intrinsically low nucleosome occupancy at sites destined for remodelling.

Low nucleosome occupancy encoded at functional binding sites

For any given transcription factor, some of its canonical target sites in the genome are occupied by a nucleosome, whereas others are not. Many of the unoccupied sites are thought to occur at random and to be functionally irrelevant^{20,41}, but the mechanism by which they are kept unoccupied is not known. An intriguing hypothesis is that genomes use their intrinsic nucleosome organization for this task by encoding stable nucleosomes over non-functional sites, thereby decreasing their accessibility to transcription factors (Fig. 4b). We tested this hypothesis by examining our predictions at binding sites for 46 transcription factors. Notably, for 17 (37%) transcription factors the predicted nucleosome occupancy at their functional and conserved DNA binding sites²⁰ was significantly lower compared with predicted occupancy at their other canonical (but presumed non-functional) sites (Fig. 4c). Only one (2%) factor exhibited significantly higher predicted occupancy at its functional binding sites. These results illustrate how the intrinsic nucleosome organization may help in directing transcription factors towards the appropriate subset of their target sites while excluding them from irrelevant sites.

Low nucleosome occupancy encoded at transcription start sites

Recent nucleosome maps indicate that nucleosomes are depleted from transcriptional start sites³¹ (TSSs), but the mechanism for this depletion is not known. For two promoter regions, this depletion was shown experimentally to be intrinsically encoded in the DNA sequence⁹. We asked whether this intrinsically encoded depletion occurs globally by examining the encoded nucleosome organization at all TSSs in yeast (Fig. 5a). We found that the most probable location for TATA elements⁴² places them in areas of the genomic sequence that remain unoccupied by nucleosomes; that is, just outside a stably positioned nucleosome (Fig. 5b). Strikingly, the location of the stably positioned nucleosome is conserved across all fungal species (Supplementary Fig. 38). We obtained all of the above results independently, applying both the chicken and yeast models to the yeast genomes. Together, these results may indicate that eukaryotic genomes direct the transcriptional machinery to functional sites by encoding unstable nucleosomes over these elements, thereby enhancing their accessibility.

Conclusions and prospects

Our results establish that nucleosome organization is encoded in eukaryotic genomes. This newly characterized genetic information occurs chromosome-wide, explains ~50% of the *in vivo* nucleosome organization, and may facilitate specific chromosome functions. The consistency between the predictions on the yeast genome using models derived independently from information concerning only yeast or chicken nucleosomes implies that the genomic signals for nucleosome positioning are strong.

Despite its successes, our approach has several limitations and represents only a first step towards understanding the DNA preferences of nucleosomes and the biological implications. First, additional experiments are needed to derive a more accurate nucleosome–DNA interaction model. Second, our representation of nucleosome–nucleosome interactions derived from a thermodynamic model does not yet account for favourable interactions⁴³, or for

the steric hindrance constraints implied by the three-dimensional nucleosome structure. Finally, we examined the intrinsic nucleosome organization without regard for the collection of DNA binding proteins that influence nucleosome positioning by competing for DNA occupancy. At equilibrium, this competition would depend on the concentrations and sequence specificities of both the DNA binding proteins and nucleosomes. The DNA binding proteins have high binding specificity but are present at low concentrations, whereas the nucleosomes have lower binding specificity but are present at high concentrations, covering 75–90% of the DNA. Thus, both are expected to make important contributions to the outcome (Supplementary Figs 39 and 40).

Overall, our results establish that genomes encode the positioning and stability of nucleosomes in regions that are critical for gene regulation and for other specific chromosome functions, and establish that this nucleosome positioning code can be successfully decoded. The genome-wide predictions of nucleosome occupancy and stability that we generated should facilitate the understanding of specific natural gene regulatory phenomena, such as the mechanism by which transcription factors bind preferentially to appropriate sites in promoters rather than to the excess of irrelevant sites in the genome. Our approach may also be useful for improving the performance of engineered transgenes. Our model and results provide a concrete framework for quantitatively integrating chromatin structure into models of gene regulation, and thus represent an essential step towards the goal of developing a quantitative, predictive understanding of transcriptional regulation in all eukaryotes.

METHODS

See Supplementary Information for a more detailed description of the methods. **Molecular biology methods.** Mononucleosomes were extracted from log-phase yeast (*Saccharomyces cerevisiae*) cells using standard methods. The DNA was extracted, and protected fragments of length ~147 bp were cloned and sequenced. An *in vitro* selection for nucleosome formation on the yeast genome was performed using purified yeast genomic DNA and substoichiometric purified histone octamer by salt gradient dialysis⁴⁴. The resulting chromatin was treated as for the *in vivo* selection. *In vitro* affinity measurements for core histone H3H4₂ tetramers were performed as described⁴⁴. *In vivo* nucleosome occupancies were measured as described⁹.

Probabilistic nucleosome–DNA interaction model. Given a collection of nucleosome DNA sequences, we aligned all sequences and their reverse complements about their centres, and associated a dinucleotide distribution with each position i , estimated from the combined dinucleotide counts at three neighbouring positions, such that the probability assigned by the model to a 147-bp sequence S is:

$$P(S) = P_1(S_1) \prod_{i=2}^{147} P_i(S_i | S_{i-1})$$

Thermodynamic model for predicting nucleosome positions genome-wide.

We used the above probabilistic nucleosome–DNA model within a statistical mechanics framework to compute the nucleosome organization intrinsic to the genomic DNA sequence. We took the partition function to be all 'legal configurations' of nucleosomes on a sequence S , where a legal configuration specifies start positions for a set of non-overlapping 147-bp nucleosomes on S , thus respecting steric hindrance effects between nucleosomes. Using our probabilistic model and an apparent nucleosome concentration parameter, we assigned a statistical weight to each configuration and used the Boltzmann distribution to compute the probability of every configuration. A dynamic programming method¹⁹ was used to efficiently compute the probability that each base pair of S starts a nucleosome or is occupied by a nucleosome.

Additional methods and URLs. For our data, model and genome-wide occupancy predictions, see <http://genie.weizmann.ac.il/pubs/nucleosomes06>. Our results are also viewable in Genomica (<http://Genomica.weizmann.ac.il>).

Received 16 March; accepted 14 June 2006.

Published online 19 July 2006.

1. Richmond, T. J. & Davey, C. A. The structure of DNA in the nucleosome core. *Nature* **423**, 145–150 (2003).
2. Satchwell, S. C., Drew, H. R. & Travers, A. A. Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.* **191**, 659–675 (1986).

3. Widom, J. Role of DNA sequence in nucleosome stability and dynamics. *Q. Rev. Biophys.* **34**, 269–324 (2001).
4. van Holde, K. E. *Chromatin* (Springer, New York, 1989).
5. Jenuwein, T. & Allis, C. D. Translating the histone code. *Science* **293**, 1074–1080 (2001).
6. Kornberg, R. D. & Lorch, Y. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* **98**, 285–294 (1999).
7. Wyrick, J. J. *et al.* Chromosomal landscape of nucleosome-dependent gene expression and silencing in yeast. *Nature* **402**, 418–421 (1999).
8. Trifonov, E. N. Sequence-dependent deformational anisotropy of chromatin DNA. *Nucleic Acids Res.* **8**, 4041–4053 (1980).
9. Sekinger, E. A., Moqtaderi, Z. & Struhl, K. Intrinsic histone–DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast. *Mol. Cell* **18**, 735–748 (2005).
10. Anderson, J. D. & Widom, J. Poly(dA–dT) promoter elements increase the equilibrium accessibility of nucleosomal DNA target sites. *Mol. Cell Biol.* **21**, 3830–3839 (2001).
11. Thåström, A. *et al.* Sequence motifs and free energies of selected natural and non-natural nucleosome positioning DNA sequences. *J. Mol. Biol.* **288**, 213–229 (1999).
12. Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425**, 737–741 (2003).
13. Cairns, B. R. Chromatin remodeling complexes: strength in diversity, precision through specialization. *Curr. Opin. Genet. Dev.* **15**, 185–190 (2005).
14. Olson, W. K., Gorin, A. A., Lu, X. J., Hock, L. M. & Zhurkin, V. B. DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. *Proc. Natl Acad. Sci. USA* **95**, 11163–11168 (1998).
15. Wang, J. P. & Widom, J. Improved alignment of nucleosome DNA sequences using a mixture model. *Nucleic Acids Res.* **33**, 6743–6755 (2005).
16. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B* **39**, 1–39 (1977).
17. Lowary, P. T. & Widom, J. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *J. Mol. Biol.* **276**, 19–42 (1998).
18. Widlund, H. R. *et al.* Identification and characterization of genomic nucleosome-positioning sequences. *J. Mol. Biol.* **267**, 807–817 (1997).
19. Rabiner, L. R. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proc. IEEE* **77**, 257–286 (1989).
20. Harbison, C. T. *et al.* Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99–104 (2004).
21. Li, S. & Smerdon, M. J. Nucleosome structure and repair of N-methylpurines in the *GAL1–10* genes of *Saccharomyces cerevisiae*. *J. Biol. Chem.* **277**, 44651–44659 (2002).
22. Weiss, K. & Simpson, R. T. High-resolution structural analysis of chromatin at specific loci: *Saccharomyces cerevisiae* silent mating type locus *HML α* . *Mol. Cell Biol.* **18**, 5392–5403 (1998).
23. Weiss, K. & Simpson, R. T. Cell type-specific chromatin organization of the region that governs directionality of yeast mating type switching. *EMBO J.* **16**, 4352–4360 (1997).
24. Moreira, J. M. & Holmberg, S. Nucleosome structure of the yeast *CHA1* promoter: analysis of activation-dependent chromatin remodeling of an RNA-polymerase-II-transcribed gene in TBP and RNA pol II mutants defective *in vivo* in response to acidic activators. *EMBO J.* **17**, 6028–6038 (1998).
25. Shimizu, M., Roth, S. Y., Szent-Gyorgyi, C. & Simpson, R. T. Nucleosomes are positioned with base pair precision adjacent to the $\alpha 2$ operator in *Saccharomyces cerevisiae*. *EMBO J.* **10**, 3033–3041 (1991).
26. Kent, N. A., Tsang, J. S., Crowther, D. J. & Mellor, J. Chromatin structure modulation in *Saccharomyces cerevisiae* by centromere and promoter factor 1. *Mol. Cell Biol.* **14**, 5229–5241 (1994).
27. Verdone, L., Camilloni, G., Di Mauro, E. & Caserta, M. Chromatin remodeling during *Saccharomyces cerevisiae* *ADH2* gene activation. *Mol. Cell Biol.* **16**, 1978–1988 (1996).
28. Almer, A., Rudolph, H., Hinnen, A. & Horz, W. Removal of positioned nucleosomes from the yeast *PHO5* promoter upon *PHO5* induction releases additional upstream activating DNA elements. *EMBO J.* **5**, 2689–2696 (1986).
29. Lee, C. K., Shibata, Y., Rao, B., Strahl, B. D. & Lieb, J. D. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nature Genet.* **36**, 900–905 (2004).
30. Bernstein, B. E., Liu, C. L., Humphrey, E. L., Perlstein, E. O. & Schreiber, S. L. Global nucleosome occupancy in yeast. *Genome Biol.* **5**, R62 (2004).
31. Yuan, G. C. *et al.* Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science* **309**, 626–630 (2005).
32. Guillemette, B. *et al.* Variant histone H2A.Z is globally localized to the promoters of inactive yeast genes and regulates nucleosome positioning. *PLoS Biol.* **3**, e384 (2005).
33. Gasch, A. P. *et al.* Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11**, 4241–4257 (2000).
34. Kim, J. & Iyer, V. R. Global role of TATA box-binding protein recruitment to promoters in mediating gene expression profiles. *Mol. Cell Biol.* **24**, 8104–8112 (2004).
35. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* **25**, 25–29 (2000).
36. Hughes, T. R. *et al.* Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126 (2000).
37. Ideker, T. *et al.* Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292**, 929–934 (2001).
38. Sudarsanam, P., Iyer, V. R., Brown, P. O. & Winston, F. Whole-genome expression analysis of *snf/swi* mutants of *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **97**, 3364–3369 (2000).
39. Casolari, J. M. *et al.* Genome-wide localization of the nuclear transport machinery couples transcriptional status and nuclear organization. *Cell* **117**, 427–439 (2004).
40. Robert, F. *et al.* Global position and recruitment of HATs and HDACs in the yeast genome. *Mol. Cell* **16**, 199–209 (2004).
41. Li, H., Rhodius, V., Gross, C. & Siggia, E. D. Identification of the binding sites of regulatory proteins in bacterial genomes. *Proc. Natl Acad. Sci. USA* **99**, 11772–11777 (2002).
42. Basehoar, A. D., Zanton, S. J. & Pugh, B. F. Identification and distinct regulation of yeast TATA box-containing genes. *Cell* **116**, 699–709 (2004).
43. Cui, Y. & Bustamante, C. Pulling a single chromatin fiber reveals the forces that maintain its higher-order structure. *Proc. Natl Acad. Sci. USA* **97**, 127–132 (2000).
44. Thåström, A., Bingham, L. M. & Widom, J. Nucleosomal locations of dominant DNA sequence motifs for histone–DNA interactions and nucleosome positioning. *J. Mol. Biol.* **338**, 695–709 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. Travers for providing the chicken nucleosome core DNA sequences; M. Kubista for providing selected mouse DNA sequences; O. Rando for providing access to their nucleosome data before publication; J. Lieb, E. Nili and P. Jones for sharing their respective unpublished data; Y. Lubling for creating the supplementary website; and H. Chang, N. Friedman, U. Gaul, A. Matouschek, B. Meyer, M. Ptashne, E. Siggia and A. Tanay for useful comments on the manuscript. E.S. was supported by a fellowship from the Center for Studies in Physics and Biology at Rockefeller University and by an NIH grant. J.W. thanks the Center for their hospitality during a sabbatical. J.-P.Z.W. acknowledges support from an NIH grant and J.W. acknowledges support from two NIH grants. E.S. is the incumbent of the Soretta and Henry Shapiro career development chair.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to E.S. (eran.segal@weizmann.ac.il) or J.W. (j-widom@northwestern.edu).

PML inhibits HIF-1 α translation and neoangiogenesis through repression of mTOR

Rosa Bernardi^{1,2}, Ilhem Guernah^{1,2}, David Jin³, Silvia Grisendi^{1,2}, Andrea Alimonti^{1,2}, Julie Teruya-Feldstein², Carlos Cordon-Cardo², M. Celeste Simon⁴, Shahin Rafii³ & Pier Paolo Pandolfi^{1,2}

Loss of the promyelocytic leukaemia (PML) tumour suppressor has been observed in several human cancers. The tumour-suppressive function of PML has been attributed to its ability to induce growth arrest, cellular senescence and apoptosis. Here we identify PML as a critical inhibitor of neoangiogenesis (the formation of new blood vessels) *in vivo*, in both ischaemic and neoplastic conditions, through the control of protein translation. We demonstrate that in hypoxic conditions PML acts as a negative regulator of the synthesis rate of hypoxia-inducible factor 1 α (HIF-1 α) by repressing mammalian target of rapamycin (mTOR). PML physically interacts with mTOR and negatively regulates its association with the small GTPase Rheb by favouring mTOR nuclear accumulation. Notably, *Pml*^{-/-} cells and tumours display higher sensitivity both *in vitro* and *in vivo* to growth inhibition by rapamycin, and lack of PML inversely correlates with phosphorylation of ribosomal protein S6 and tumour angiogenesis in mouse and human tumours. Thus, our findings identify PML as a novel suppressor of mTOR and neoangiogenesis.

The 'angiogenic switch' is essential in tumorigenesis¹. Neoangiogenesis often occurs as a response to intra-tumoural hypoxia, which leads to upregulation of the transcription factor HIF-1 α and expression of its target gene *VEGF* (ref. 2). Accordingly, intra-tumoural hypoxia, overexpression of HIF-1 α , and increased microvascular density associate with tumour progression and poor prognosis³.

The promyelocytic leukaemia gene *PML* was identified at the breakpoint of the t(15;17) translocation of acute promyelocytic leukaemia⁴⁻⁷. In addition to acting as a bona fide leukaemia suppressor in acute promyelocytic leukaemia⁸, the relevance of PML as a tumour suppressor has also been extended beyond its involvement in leukaemogenesis. Forced PML expression induces growth inhibition, apoptosis and cellular senescence depending on the cellular context⁹, and *Pml* inactivation in mice results in susceptibility to cancer¹⁰⁻¹². Notably, PML is lost in a vast variety of human tumours¹³⁻¹⁶.

So far, the tumour-suppressive role of PML has been attributed to its ability to regulate tumour-suppressive transcription factors such as p53, pRb and SMAD^{9,17,18}. Here we show that PML is a critical regulator of angiogenesis in ischaemia and in tumours through inhibition of HIF-1 α translation. We implicate PML in the modulation of a novel mechanism of mTOR inhibition upon hypoxia, its nuclear accumulation, and we demonstrate that PML loss sensitizes tumours to inhibitors of mTOR and angiogenesis, findings that bear important therapeutic implications.

PML opposes revascularization upon ischaemia

PML loss and neoangiogenesis are associated with tumour progression^{3,16}. To test whether *Pml* loss promotes angiogenesis, wild-type and *Pml*^{-/-} mice were subjected to ligation of the femoral vessels, a procedure that causes muscle ischaemia followed by

revascularization¹⁹. Blood perfusion recovery was measured as blood flow ratio of ischaemic versus non-ischaemic hindlimb over a 2-week time course. Notably, average blood flow rates in *Pml*^{-/-} mice were significantly higher than those in wild-type animals, indicating that *Pml*^{-/-} mice recovered faster from ischaemia (Fig. 1a).

Ischaemic revascularization and tumour angiogenesis are supported by the mobilization of endothelial progenitor cells from the bone marrow in response to circulating cytokines produced by ischaemic tissues²⁰. Their incorporation into newly formed blood vessels is essential for the formation of a functional vasculature²¹. Circulating endothelial progenitor cells (CD11b⁻VEGFR2⁺ cells²²) were more abundant in *Pml*^{-/-} mice than in wild-type animals after surgery (Fig. 1b).

Next, the extent of revascularization in the ischaemic muscles was measured by immunostaining with CD31 antibodies. The amount of endothelial cells per muscle fibre was significantly higher in *Pml*^{-/-} mice (Fig. 1c). Taken together, these experiments show that absence of Pml greatly facilitates the formation of new blood vessels in response to ischaemic injury, indicating that *Pml* has an important physiological role in suppressing revascularization upon ischaemia.

Revascularization and angiogenesis are mediated by the production of pro-angiogenic factors such as the cytokine VEGF, a target gene of the transcription factor HIF-1 and a major regulator of angiogenesis². The hypoxia-regulated subunit of Hif-1, Hif-1 α , markedly accumulates in ischaemic muscles upon femoral artery ligation, peaking at day 3 after surgery²³. We analysed plasma levels of VEGF and the expression of Hif-1 α in the adductor and gastrocnemius muscles of wild-type and *Pml*^{-/-} mice. Both plasma VEGF and muscle Hif-1 α accumulation were higher in *Pml*^{-/-} mice (Fig. 1d, e). Considerable accumulation of Hif-1 α in wild-type mice could be

¹Cancer Biology and Genetics Program, ²Department of Pathology, Memorial Sloan-Kettering Cancer Center, Sloan-Kettering Institute, 1275 York Avenue, New York, New York 10021, USA. ³Howard Hughes Medical Institute, Department of Genetic Medicine, Division of Hematology-Oncology, Weill-Cornell Medical College, New York, New York 10021, USA. ⁴Howard Hughes Medical Institute, Department of Cell & Developmental Biology and Abramson Family Cancer Research Institute, University of Pennsylvania Cancer Center, Philadelphia, Pennsylvania 19104, USA.

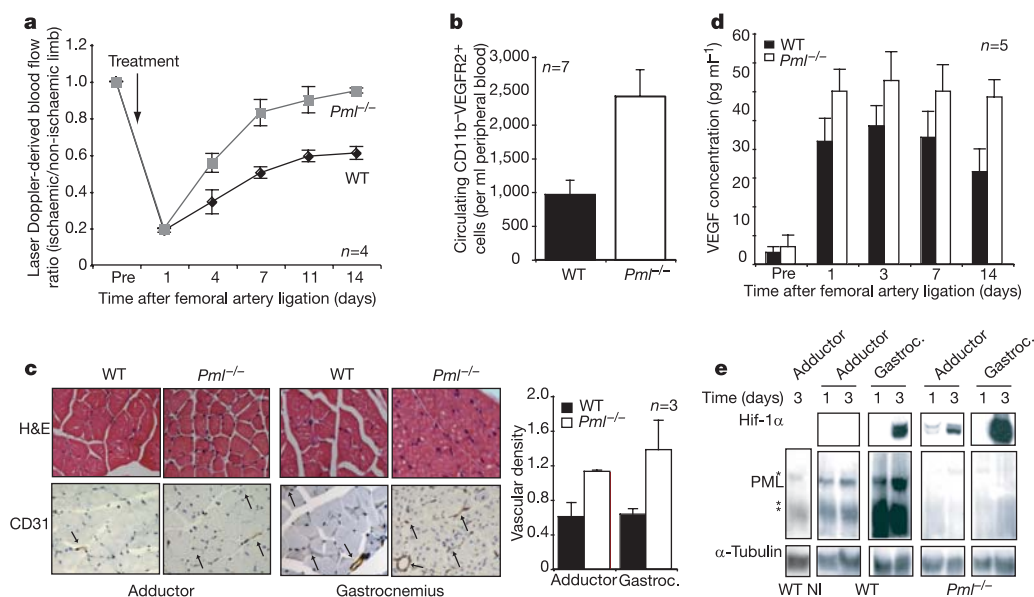


Figure 1 | *Pml*^{-/-} mice have accelerated revascularization in response to ischaemia. **a**, Blood flow ratio of ischaemic/non-ischaemic limb in wild-type and *Pml*^{-/-} mice. One representative experiment is shown out of three with similar results. *P* < 0.02 at all time points. Pre, measurement before surgery. **b**, Flow cytometry from peripheral blood of wild-type and *Pml*^{-/-} mice (day 7 after surgery; average of three independent experiments; *P* < 0.001). **c**, Haematoxylin and eosin (H&E) staining and

immunohistochemistry of CD31-positive cells (arrows) in muscles of wild-type and *Pml*^{-/-} mice at day 15 after surgery. The graph on the right shows quantification of endothelial cells per muscle fibre. Ten fields per muscle were counted. **d**, VEGF-A ELISA from plasma of wild-type and *Pml*^{-/-} mice. Error bars in **a–d** represent s.d. **e**, Expression of Hif-1α, Pml (asterisks indicate different isoforms) and α-tubulin in muscles of wild-type and *Pml*^{-/-} mice. Non-ischaemic (NI) wild-type muscle serves as control.

detected only in the gastrocnemius muscle (the most distal to the site of artery ligation), whereas in *Pml*^{-/-} mice, Hif-1α levels were higher in the gastrocnemius and were also detectable in the adductor muscle (Fig. 1e). Notably, in wild-type animals Pml protein levels increased upon ischaemia (Fig. 1e). These data indicate that the accelerated revascularization in *Pml*^{-/-} mice might be mediated by

increased local response to hypoxia, prompting us to investigate whether Pml directly controls the transcription factor Hif-1α.

Pml is a negative regulator of Hif-1α

To assess whether PML directly regulates HIF-1α activity, we performed luciferase reporter assays. Co-transfection of PML with

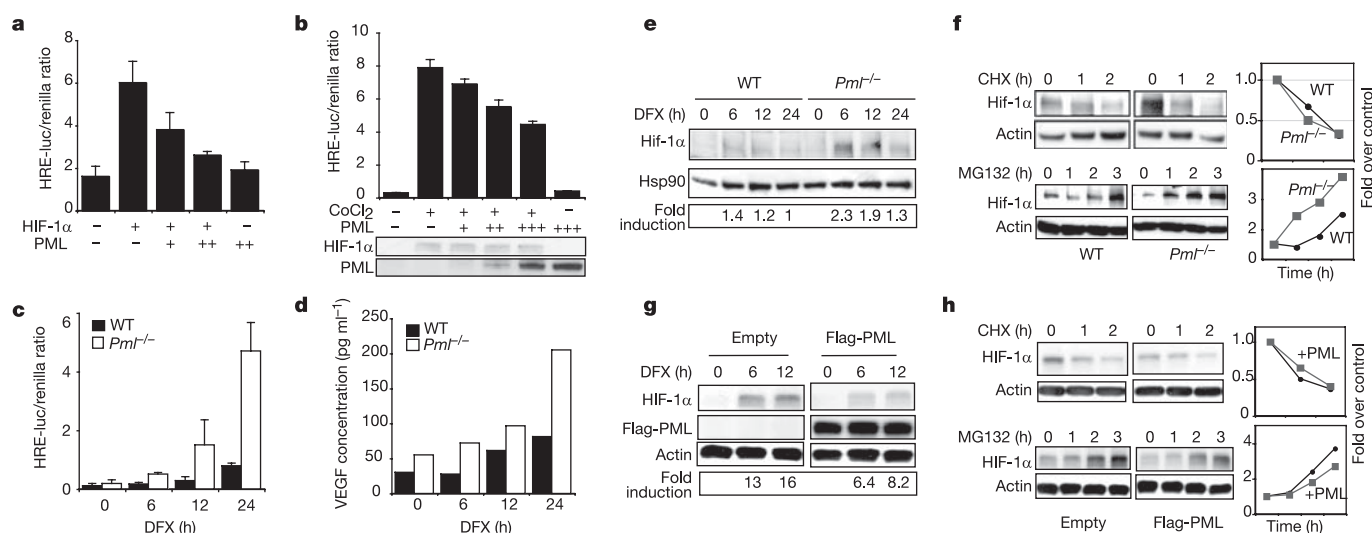


Figure 2 | Pml controls Hif-1α activity and accumulation rate in hypoxic conditions. **a–c**, HIF-1α-transactivation assays with hypoxia-response element (HRE) luciferase. **a**, H1299 cells transfected with HIF-1α and increasing concentrations of PML. **b**, H1299 cells transfected with PML and treated with CoCl₂. Western blots at the bottom show HIF-1α and PML expression. **c**, Wild-type and *Pml*^{-/-} MEFs were treated with DFX. Results in **a–c** are presented as mean ± s.d. from one experiment performed in triplicate. Experiments were repeated three times. **d**, VEGF-A concentration in media from wild-type and *Pml*^{-/-} MEFs treated with DFX. **e**, Western blot of Hif-1α in wild-type and *Pml*^{-/-} MEFs treated

with DFX. Numbers express Hif-1α levels upon normalization for Hsp90 expressed as fold induction over control. **f**, Western blot of Hif-1α in wild-type and *Pml*^{-/-} MEFs treated with DFX for 6 h followed by treatment with cycloheximide (CHX) or MG132. Graphs express Hif-1α levels upon normalization for β-actin as fold over cells treated with DFX. **g**, Western blot of HIF-1α and Flag-PML in PC-3 cells stably transfected with PML and treated with DFX. HIF-1α levels are expressed as in **e**. **h**, Western blot of HIF-1α in PC-3 cells treated as in **f**. Graphs express HIF-1α levels as in **f**. **d–f** are representative experiments out of three independent experiments with similar results.

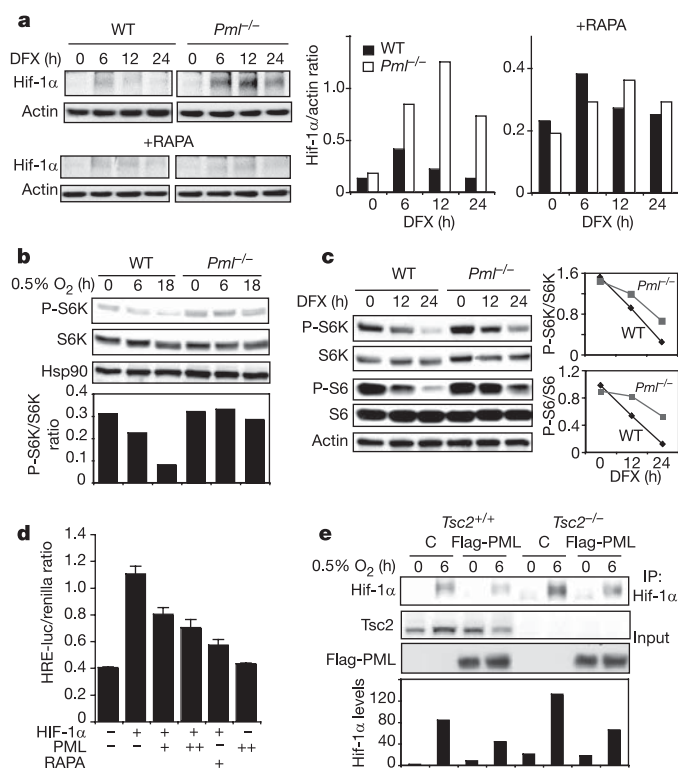


Figure 3 | PML regulates mTOR in response to hypoxia. **a**, Western blot of Hif-1 α in wild-type and *Pml*^{-/-} MEFs treated with DFX and co-treated with rapamycin (RAPA) for 24 h. Graphs represent Hif-1 α levels upon normalization for β -actin. **b**, **c**, Western blot of phosphorylated S6K (P-S6K), S6K, P-S6 and S6 in wild-type and *Pml*^{-/-} MEFs treated with 0.5% O₂ (**b**) and DFX (**c**). Graphs represent P-S6K and P-S6 levels normalized for total protein levels and β -actin. **d**, HIF-1 α transactivation assay in H1299 cells transfected with PML or treated with rapamycin for 24 h. Results are presented as mean $\pm$ s.d. from one experiment performed in triplicate. **e**, Immunoprecipitation–western blot of Hif-1 α in *Tsc2*^{+/+} and *Tsc2*^{-/-} cells stably transfected with empty vector (C) or PML in response to hypoxia. Expression levels of Tsc2 and Flag-PML in the inputs are shown in the lower panels. The bar graph at the bottom represents Hif-1 α levels. All experiments were repeated at least three times with similar results.

HIF-1 α revealed that PML represses HIF-1 α -mediated transactivation in a dose-dependent manner (Fig. 2a). Moreover, PML inhibits the transcriptional activity of endogenous HIF-1 α in cells treated with hypoxia-mimetic agents CoCl₂ and desferrioxamine (DFX), or under conditions of hypoxia (Fig. 2b, data not shown and Supplementary Fig. 1a). Conversely, Hif-1 α transactivation was higher in *Pml*^{-/-} compared with wild-type mouse embryonic fibroblasts (MEFs) treated with hypoxia-mimetic drugs or under conditions of hypoxia (Fig. 2c and Supplementary Fig. 1b); this was also the case for the endogenous Hif-1 α targets VEGF and Glut-1 (Fig. 2d and Supplementary Fig. 1c, d).

We next asked whether increased Hif-1 α activity in *Pml*^{-/-} cells results from increased protein accumulation. Consistent with what was observed in ischaemic muscles, western blot analysis of Hif-1 α in wild-type and *Pml*^{-/-} MEFs treated with hypoxia-mimetic agents or under conditions of hypoxia revealed that Hif-1 α protein levels were higher in the absence of *Pml* (Fig. 2e and Supplementary Fig. 1e). Real-time polymerase chain reaction with reverse transcription (RT-PCR) did not reveal any induction of Hif-1 α mRNA in wild-type or *Pml*^{-/-} MEFs treated with hypoxia-mimetic agents (not shown), indicating that *Pml* controls Hif-1 α accumulation at the post-transcriptional level.

Hif-1 α protein levels are tightly controlled by oxygen concentration^{24,25}. In normoxia, Hif-1 α is constantly hydroxylated and targeted for ubiquitination by an E3-ubiquitin ligase containing

the Von Hippel–Lindau tumour suppressor pVHL^{24–26}, whereas Hif-1 α degradation is inhibited in hypoxia and by treatment with hypoxia-mimetic agents. To test whether Hif-1 α degradation was inhibited in *Pml*^{-/-} cells, we calculated Hif-1 α half-life upon treatment with DFX (Fig. 2f, upper panels). We did not detect any significant difference between wild-type and *Pml*^{-/-} MEFs. Notably, however, Hif-1 α protein synthesis rates, measured by blocking degradation with the proteasome inhibitor MG132, were higher in *Pml*^{-/-} than in wild-type MEFs upon treatment with DFX and hypoxia (Fig. 2f, lower panels and Supplementary Fig. 1f).

To corroborate these findings, we stably transfected tumour cell lines with PML and studied HIF-1 α accumulation. Overexpression of PML caused reduced accumulation of HIF-1 α upon treatment with hypoxia-mimetic agents and hypoxia in PC-3 (Fig. 2g and Supplementary Fig. 2c), MCF-7 (Supplementary Fig. 2a) and SW-620 (Supplementary Fig. 2b) cells. Furthermore, in PC-3 cells treated with DFX and hypoxia, HIF-1 α half-life was not affected by PML overexpression, whereas HIF-1 α accumulation rate was inhibited (Fig. 2h and Supplementary Fig. 2d).

To exclude further that PML promotes HIF-1 α degradation, we made use of renal carcinoma cells that harbour a *VHL* mutation and thus express constitutively high levels of HIF-1 α . PML caused a decrease in the levels of HIF-1 α that could not be rescued by treatment with MG132 (Supplementary Fig. 2e), demonstrating that PML regulates HIF-1 α levels in a pVHL-independent manner.

Taken together, these data reveal that PML acts as a negative regulator of HIF-1 α through regulation of its synthesis rate.

PML is a negative regulator of mTOR

Recent data have shown that hypoxia leads to inhibition of mTOR²⁷ through HIF-dependent and -independent mechanisms that, at least in part, depend on TSC2 (refs 28, 29). In *Tsc2*^{-/-} MEFs, Hif-1 α accumulates at higher levels compared with wild-type cells under conditions of hypoxia, and this can be prevented by treatment with rapamycin²⁸. As in *Tsc2*^{-/-} MEFs, rapamycin abolished the difference in Hif-1 α accumulation between wild-type and *Pml*^{-/-} MEFs (Fig. 3a and Supplementary Fig. 3a), indicating that higher accumulation levels of Hif-1 α in *Pml*^{-/-} MEFs could be due to increased mTOR activity. We therefore analysed the phosphorylation of S6 kinase (S6K) on amino acid T389, the major rapamycin-sensitive site³⁰. Levels of S6K phosphorylation diminished less in *Pml*^{-/-} than in wild-type MEFs upon hypoxia, DFX and CoCl₂ treatment (Fig. 3b, c and Supplementary Fig. 3b). The same was true for the phosphorylation of S6 (S235/236) (Fig. 3c). Notably, in *Pml*^{-/-} MEFs the decrease in phosphorylated S6K levels was less compared with wild type also upon serum starvation, and the same was evident in 293 cells upon PML knockdown (Supplementary Fig. 3c, d). Moreover, an increase in Akt phosphorylation levels on S473 was noted in wild-type but not *Pml*^{-/-} MEFs treated with DFX, consistent with suppression of the negative feedback³¹ (Supplementary Fig. 3e). Interestingly, rapamycin, like PML expression, inhibited HIF-1 α -mediated transactivation even when Hif-1 α was expressed as a cDNA lacking 5' untranslated regions (UTRs) (Fig. 3d).

Thus, *Pml*^{-/-} MEFs behave similarly to *Tsc2*^{-/-} cells in hypoxia, which prompted us to test whether PML inhibits Hif-1 α accumulation in a Tsc2-dependent manner. Notably, stable overexpression of PML decreased Hif-1 α levels in both wild-type and *Tsc2*^{-/-} MEFs (Fig. 3e). Part of the Tsc2-mediated inhibition of mTOR in hypoxia has been attributed to transcriptional upregulation of Redd1 (ref. 28). No significant difference in the induction of Redd1 was noted between wild-type and *Pml*^{-/-} MEFs (Supplementary Fig. 3f). Additionally, short-interfering RNA (siRNA)-mediated knockdown of PML caused increased HIF-1 α levels in PC-3 cells treated with DFX; however, REDD1 induction was similar (or perhaps slightly higher) compared to control siRNA cells (Supplementary Fig. 3g).

Taken together, these results indicate that PML controls HIF-1 α

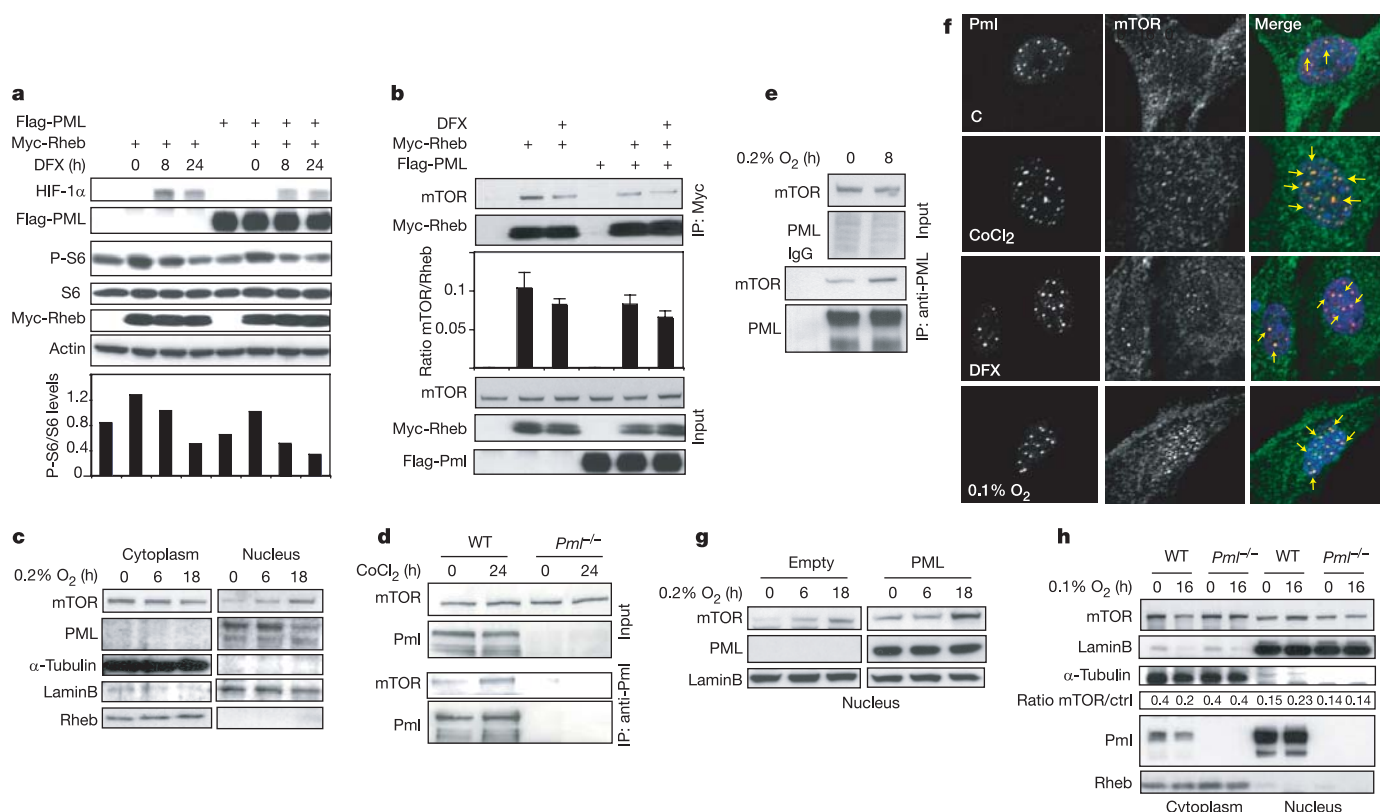


Figure 4 | PML inhibits mTOR activity through physical interaction and co-localization in the nucleus. **a**, Western blot analysis of HIF-1 α , PML, P-S6, S6, Rheb and actin in 293 cells transfected with Flag-PML and Myc-Rheb and treated with DFX. Graph represents P-S6 levels normalized for total protein levels and β -actin. **b**, Immunoprecipitation with anti-Myc antibody from 293 cells transfected with Myc-Rheb with or without Flag-PML and treated with DFX for 8 h. Top panels: Rheb and mTOR in immunoprecipitation. The bar graph shows mTOR/Rheb ratios in immunoprecipitations from three different experiments. Error bars represent s.d. Bottom panels: input levels of mTOR, Myc-Rheb and Flag-PML. **c**, Nuclear and cytoplasmic localization of mTOR, PML and Rheb in 293 cells treated with 0.2% O₂. α -Tubulin and laminB represent loading

controls. Note that PML levels decrease at late time points in hypoxia. **d**, **e**, Co-immunoprecipitation of Pml and mTOR in MEFs treated with CoCl₂ (**d**) and 293 cells treated with 0.2% O₂ (**e**). **f**, Representative confocal images of co-localization between Pml and mTOR (Cell Signaling antibody) in wild-type control MEFs and 24 h after CoCl₂, DFX and 0.1% O₂ treatment. Arrows indicate co-localization of Pml and mTOR. **g**, Nuclear accumulation of mTOR in 293 cells treated with 0.2% O₂ upon PML overexpression. **h**, Nuclear and cytoplasmic localization of mTOR, Pml and Rheb in wild-type and *Pml*^{-/-} MEFs upon treatment with 0.1% O₂. Numbers indicate ratios of mTOR over nuclear and cytoplasmic controls. All experiments were repeated at least three times with similar results.

accumulation in hypoxic conditions in an mTOR-dependent, TSC2/REDD1-independent manner.

PML inhibits Rheb–mTOR association

We next assessed whether PML directly modulates mTOR activity. PML overexpression in 293 cells caused a dose-dependent inhibition in the phosphorylation of S6K, S6 and 4EBP1 (Supplementary Fig. 4a). Notably, PML also inhibited the activity of wild-type S6K but not of a rapamycin-insensitive S6K mutant (Supplementary Fig. 4b).

Epistatic and biochemical studies have recently established that mTOR activity is stimulated by the small GTPase Rheb, which is opposed by the TSC1–TSC2 complex acting as a GTPase-activating protein (GAP)³². Rheb overexpression in 293 cells caused an increase in phosphorylated S6 levels that was inhibited by DFX and hypoxia treatment and further inhibited by PML co-expression (Fig. 4a and Supplementary Fig. 4c). Furthermore, HIF-1 α accumulation was inhibited by transient expression of PML also in the presence of Rheb (Fig. 4a).

It was recently shown that Rheb association with mTOR is inhibited by amino acid or leucine withdrawal, conditions leading to mTOR inhibition³³. We investigated whether treatment with hypoxia and hypoxia-mimetic agents inhibits Rheb–mTOR association. Indeed, the amount of mTOR co-immunoprecipitated with Myc-Rheb diminished after DFX treatment and hypoxia, and PML overexpression increased the dissociation (Fig. 4b and Supplementary Fig. 4d).

mTOR interacts with PML upon hypoxia

To determine the mechanism causing Rheb–mTOR dissociation, and the role of PML in this process, we analysed the subcellular localization of these proteins. We found that mTOR accumulates in the nucleus upon hypoxia (Fig. 4c) and DFX treatment (Supplementary Fig. 4e), whereas Rheb remains exclusively cytoplasmic (both endogenous, Fig. 4c, and overexpressed, not shown). This is consistent with previous findings showing that mTOR shuttles between the cytoplasm and the nucleus^{34,35}, and that conditions that inhibit its activity, such as rapamycin and γ -radiation, promote its nuclear accumulation³⁶. Notably, PML and mTOR were found to interact physically (Fig. 4d, e) and co-localize in PML nuclear bodies (PML-NBs³⁷, as assessed by three different mTOR antibodies; Fig. 4f, Supplementary Fig. 4f and data not shown), especially upon hypoxia and treatment with hypoxia-mimetic agents. Furthermore, overexpression of PML enhanced the nuclear accumulation of mTOR in hypoxia (Fig. 4g) whereas *Pml*^{-/-} MEFs showed a defect in mTOR nuclear accumulation (Fig. 4h). Taken together, these data show that in hypoxic conditions PML inhibits mTOR association with Rheb through physical interaction and association in the nucleus.

PML antagonizes tumour angiogenesis

To study the role of Pml in regulating tumour angiogenesis, tumorigenesis assays were performed with transformed MEFs. Despite showing similar proliferation rates and colony forming ability

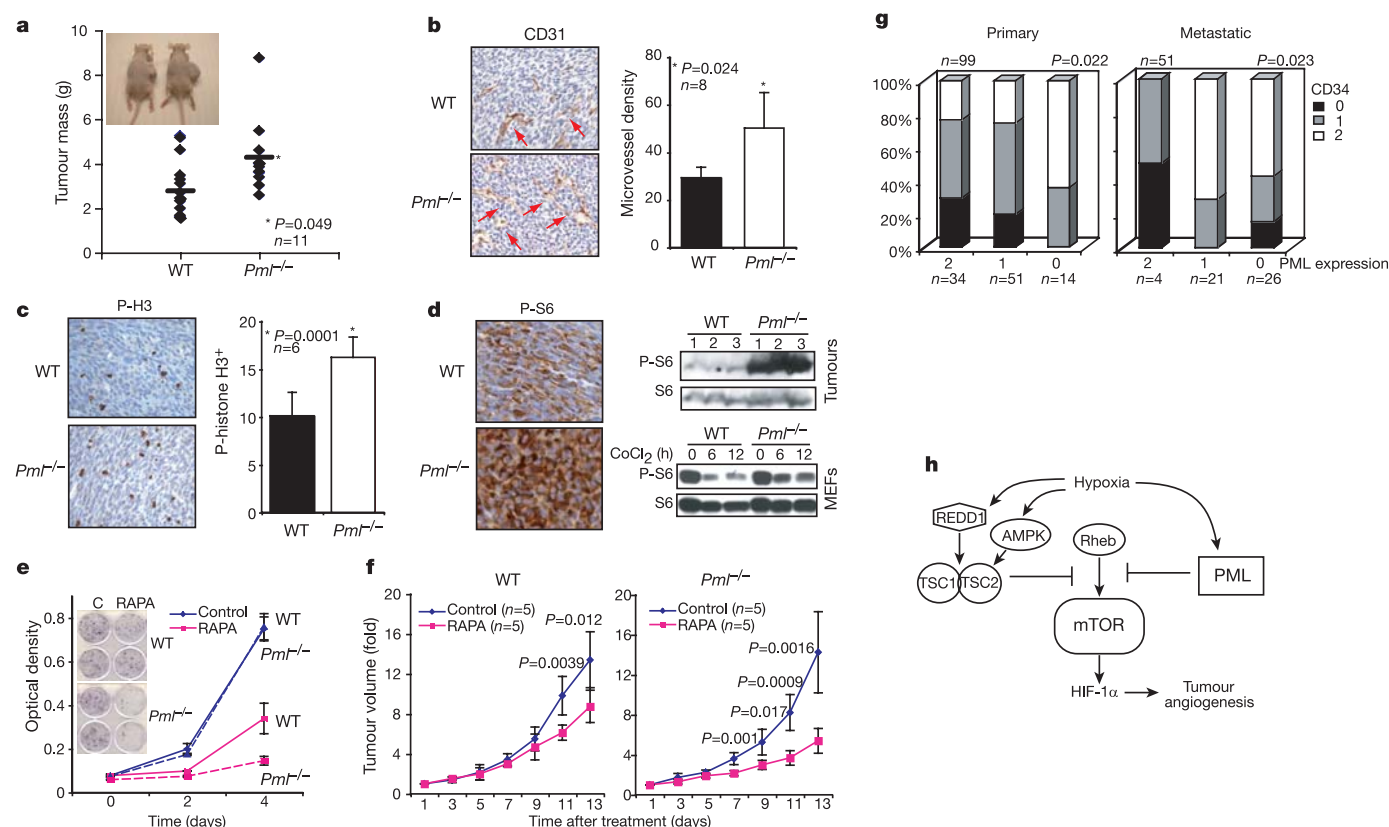


Figure 5 | *Pml*^{-/-} tumours have increased angiogenesis, phosphorylation of S6 and sensitivity to rapamycin. **a**, Tumour weight of wild-type and *Pml*^{-/-} transformed MEFs (Ras^{V12} + E1A) 18 days after injection in nude mice. The graph combines three independent experiments with three different sets of MEFs. **b**, CD31 immunostaining of wild-type and *Pml*^{-/-} tumours. Arrows indicate microvessels and the graph shows average number of microvessels per ×20 field. **c**, Phosphorylated H3 immunostaining of wild-type and *Pml*^{-/-} tumours. The graph represents the number of phosphorylated H3-positive cells per ×40 field. For graphs in **b**, **c**, ten fields per tumour were counted. Results are presented as mean ± s.d. **d**, P-S6 immunostaining of wild-type and *Pml*^{-/-} tumours of similar sizes. Western blots on the right show P-S6 and S6 in extracts from tumours (top) and from transformed MEFs upon CoCl₂ treatment (bottom). **e**, Representative

growth curve, out of three independently performed, of transformed wild-type and *Pml*^{-/-} MEFs untreated or treated with rapamycin. Results are presented as mean ± s.d. from one experiment performed in triplicate. The inset shows representative examples of wells from 4-day cultures upon staining with crystal violet. **f**, Tumour growth in nude mice injected with wild-type (left) or *Pml*^{-/-} (right) transformed MEFs treated with rapamycin or vehicle. Results are presented as mean ± s.d. from one out of two independent experiments with similar results. Statistical significance at each time point was determined by Student's *t*-test. **g**, Bar graph showing correlation of high microvessel density (measured as CD34 immunohistochemistry) with low PML expression in primary and metastatic prostate tumours. Statistical significance was determined by a Pearson chi-squared test. **h**, Mechanisms of mTOR inhibition in hypoxia.

in vitro (Supplementary Fig. 5a, b), *Pml*^{-/-} MEFs generated larger fibrosarcomas than wild-type MEFs when injected in nude mice (Fig. 5a). This indicates that *Pml*^{-/-} MEFs exhibit *in vivo* tumour-promoting properties unnecessary *in vitro*. Indeed, *Pml*^{-/-} tumours had greater microvessel density (Fig. 5b) and a high number of haemorrhagic lesions (tumour blood vessels are often malformed, leading to haemorrhages; Supplementary Fig. 5d), as well as a higher proliferation index (Fig. 5c). Notably, *Pml*^{-/-} tumours displayed greater microvessel density also at early time points after injection of MEFs, at a stage where they do not yet exhibit greater proliferation (Supplementary Fig. 5e, f). Additionally, a marked increase in phosphorylated S6 staining was detected both by immunohistochemistry and western blots of extracts from *Pml*^{-/-} tumours (Fig. 5d). This can be explained as a result of intra-tumoural hypoxia and higher mTOR activity in the absence of Pml, and is consistent with higher levels of phosphorylated S6 in transformed *Pml*^{-/-} MEFs before injection (Fig. 5d, MEFs).

On the basis of these findings, we studied the sensitivity of *Pml*^{-/-} transformed cells to mTOR and angiogenesis inhibitors *in vitro* and *in vivo*. Remarkably, *Pml*^{-/-} transformed cells displayed higher sensitivity to rapamycin (Fig. 5e). In addition, *Pml*^{-/-} tumours were also significantly more sensitive to systemic treatment with rapamycin (Fig. 5f). Moreover, *Pml*^{-/-} tumours were more sensitive

to a monoclonal antibody against VE-cadherin that inhibits tumour growth by blocking tumour angiogenesis^{38,39} (Supplementary Fig. 5g). Hence, in a genetically predetermined background, tumours lacking Pml exhibit increased microvessel density, increased proliferation and higher phosphorylated S6 staining, as well as increased sensitivity to rapamycin and angiogenesis blockade, consistent with aberrant mTOR signalling and angiogenesis.

To validate the relevance of our findings to human cancer, we analysed the expression levels of PML in correlation with tumour angiogenesis in human prostate cancer. Tissue microarray (TMA) analysis was performed on primary and metastatic prostate cancer. In primary tumours (low and high grade: 6–7 and 8–9 Gleason score, respectively), PML loss correlated with disease progression, in agreement with our previous analysis¹⁶ (Supplementary Fig. 5h). Tumour angiogenesis also correlated with disease progression (Supplementary Fig. 5h). Notably, in both primary and metastatic prostate cancer, PML loss correlated with increased microvessel density (Fig. 5g).

Discussion

In this study we find that PML is a critical regulator of angiogenesis in both neoplastic and ischaemic conditions. Owing to the relevance of neoangiogenesis in these pathological processes, the identification of novel molecular targets involved in its regulation is important for the

development of innovative therapeutic strategies.

Mechanistically, we identify PML as a repressor of mTOR and HIF-1 α translation and we determine a novel process of mTOR inhibition under hypoxia, which acts by promoting its nuclear accumulation and reducing its interaction with the small cytoplasmic GTPase Rheb. By integrating our data with earlier findings^{28,29}, we propose a model by which inhibition of mTOR in hypoxic conditions occurs as a result of multiple converging pathways (Fig. 5h): (1) REDD1 upregulation and AMPK activation promote mTOR inhibition through activation of the TSC1–TSC2 tumour suppressor complex^{28,29}; (2) mTOR accumulates in the nucleus through the action of the PML tumour suppressor.

Our findings bear important prognostic and therapeutic implications for the treatment of both ischaemic conditions and cancer. Analysis of PML expression levels in human cancer could be used as a biological marker to predict efficacy of treatment with mTOR or angiogenesis inhibitors. From a therapeutic standpoint, PML down-regulation could be an effective endpoint in conditions of tissue ischaemia, as mice lacking *Pml* show increased accumulation of Hif-1 α and increased revascularization upon ischaemic insult. Thus, well-tolerated drugs that are known to reduce PML protein levels, such as low doses of As₂O₃ (ref. 40), could be beneficial in the treatment of acute ischaemic conditions. Conversely, drugs that are known to upregulate PML protein and mRNA levels, such as proteasome inhibitors¹⁶ and interferons^{41,42}, could be used for cancer therapy in combination with angiogenesis inhibitors² or mTOR inhibitors^{43,44}.

METHODS

Cell culture and mice. MEFs were prepared as described⁴⁵. *Tsc2*^{-/-} *p53*^{-/-} and *Tsc2*^{+/-} *p53*^{-/-} MEFs were from D. Kwiatkowski. H1299, RCC-4, PC-3, MCF-7, SW-620 and HEK-293 (293) cells were from ATCC. For *in vivo* studies, 5×10^6 Ras^{V12} + E1A transformed MEFs were injected subcutaneously into the flank of 6-week-old athymic nude mice (NCRNU-M, Taconic Farms, Inc.). RAD001 (everolimus: a derivative of rapamycin) was administered by oral gavage daily at 10 mg⁻¹ kg⁻¹ day⁻¹ starting when tumours reached 150 mm³.

Hindlimb ischaemia and flow cytometry. Ligation and resection of femoral vessels is described in Supplementary Methods. Rate of blood flow was measured with a digital laser Doppler flowmeter (Model ALF21, Advance Co.). For flow cytometry, circulating mononuclear cells were stained with phycoerythrin-conjugated CD11b (BD PharMingen) and FITC-conjugated VEGFR2 (provided by ImClone Systems). Gating controls were set up with appropriate IgG isotypes.

Western blotting, cell fractionations and immunoprecipitation. Western blot analysis was performed as published⁴⁵. For a list of antibodies used, see Supplementary Methods. For co-immunoprecipitation, cells were lysed in CHAPS buffer⁴⁶. Lysates were pre-cleared with protein-G Sepharose beads (Amersham). Immunoprecipitations with anti-Myc (Cell Signaling) and anti-PML (PG-M3, Santa Cruz, or S36 from S. Lowe) antibodies were performed overnight at 4 °C. Nuclear–cytoplasmic fractionation was performed as described³⁴.

Immunofluorescence and immunohistochemistry. Immunofluorescence was performed as described⁴⁵. Antibodies for mTOR were from Cell Signaling, Upstate Biotechnologies and Sigma. For immunohistochemistry, tumours were fixed in 4% paraformaldehyde and embedded in paraffin. Five-micrometre-thick sections were stained with haematoxylin and eosin. Immunohistochemical detection was performed as described⁴⁷ with the following antibodies: CD31 PECAM-1 (Santa Cruz); phospho-histone H3 (P-H3, Cell Signaling); P-S6 (Cell Signaling).

Received 8 March; accepted 30 June 2006.

1. Hanahan, D. & Folkman, J. Patterns and emerging mechanisms of the angiogenic switch during tumorigenesis. *Cell* **86**, 353–364 (1996).
2. Ferrara, N. Vascular endothelial growth factor: basic science and clinical progress. *Endocr. Rev.* **25**, 581–611 (2004).
3. Semenza, G. L. HIF-1 and tumor progression: pathophysiology and therapeutics. *Trends Mol. Med.* **8**, S62–S67 (2002).
4. de The, H. *et al.* The PML-RAR α fusion mRNA generated by the t(15;17) translocation in acute promyelocytic leukemia encodes a functionally altered RAR. *Cell* **66**, 675–684 (1991).
5. Goddard, A. D., Borrow, J., Freemont, P. S. & Solomon, E. Characterization of a zinc finger gene disrupted by the t(15;17) in acute promyelocytic leukemia. *Science* **254**, 1371–1374 (1991).

6. Kakizuka, A. *et al.* Chromosomal translocation t(15;17) in human acute promyelocytic leukemia fuses RAR α with a novel putative transcription factor, PML. *Cell* **66**, 663–674 (1991).
7. Pandolfi, P. P. *et al.* Structure and origin of the acute promyelocytic leukemia myl/RAR α cDNA and characterization of its retinoid-binding and transactivation properties. *Oncogene* **6**, 1285–1292 (1991).
8. Rego, E. M. *et al.* Role of promyelocytic leukemia (PML) protein in tumor suppression. *J. Exp. Med.* **193**, 521–529 (2001).
9. Salomoni, P. & Pandolfi, P. P. The role of PML in tumor suppression. *Cell* **108**, 165–170 (2002).
10. Wang, Z. G. *et al.* PML is essential for multiple apoptotic pathways. *Nature Genet.* **20**, 266–272 (1998).
11. Wang, Z. G. *et al.* Role of PML in cell growth and the retinoic acid pathway. *Science* **279**, 1547–1551 (1998).
12. Trotman, L. C. *et al.* Identification of a tumour suppressor network opposing nuclear Akt function. *Nature* **441**, 523–527 (2006).
13. Koken, M. H. *et al.* The PML growth-suppressor has an altered expression in human oncogenesis. *Oncogene* **10**, 1315–1324 (1995).
14. Gambacorta, M. *et al.* Heterogeneous nuclear expression of the promyelocytic leukemia (PML) protein in normal and neoplastic human tissues. *Am. J. Pathol.* **149**, 2023–2035 (1996).
15. Zhang, P. *et al.* Lack of expression for the suppressor PML in human small cell lung carcinoma. *Int. J. Cancer* **85**, 599–605 (2000).
16. Gurrieri, C. *et al.* Loss of the tumor suppressor PML in human cancers of multiple histologic origins. *J. Natl Cancer Inst.* **96**, 269–279 (2004).
17. Lin, H. K., Bergmann, S. & Pandolfi, P. P. Cytoplasmic PML function in TGF- β signalling. *Nature* **431**, 205–211 (2004).
18. Takahashi, Y., Lallemand-Breitenbach, V., Zhu, J. & de The, H. PML nuclear bodies and apoptosis. *Oncogene* **23**, 2819–2824 (2004).
19. Marti, H. H. & Risau, W. Angiogenesis in ischemic disease. *Thromb. Haemost.* **82** (suppl. 1), 44–52 (1999).
20. Rafii, S. & Lyden, D. Therapeutic stem and progenitor cell transplantation for organ vascularization and regeneration. *Nature Med.* **9**, 702–712 (2003).
21. Rabbany, S. Y., Heissig, B., Hattori, K. & Rafii, S. Molecular pathways regulating mobilization of marrow-derived stem cells for tissue revascularization. *Trends Mol. Med.* **9**, 109–117 (2003).
22. Lyden, D. *et al.* Impaired recruitment of bone-marrow-derived endothelial and hematopoietic precursor cells blocks tumor angiogenesis and growth. *Nature Med.* **7**, 1194–1201 (2001).
23. Amano, K. *et al.* Mechanism for IL-1 β -mediated neovascularization unmasked by IL-1 β knock-out mice. *J. Mol. Cell. Cardiol.* **36**, 469–480 (2004).
24. Semenza, G. L. Targeting HIF-1 for cancer therapy. *Nature Rev. Cancer* **3**, 721–732 (2003).
25. Kaelin, W. G. Jr. How oxygen makes its presence felt. *Genes Dev.* **16**, 1441–1445 (2002).
26. Brahimi-Horn, C. & Pouyssegur, J. When hypoxia signalling meets the ubiquitin-proteasomal pathway, new targets for cancer therapy. *Crit. Rev. Oncol. Hematol.* **53**, 115–123 (2005).
27. Arsham, A. M., Howell, J. J. & Simon, M. C. A novel hypoxia-inducible factor-independent hypoxic response regulating mammalian target of rapamycin and its targets. *J. Biol. Chem.* **278**, 29655–29660 (2003).
28. Brugarolas, J. *et al.* Regulation of mTOR function in response to hypoxia by REDD1 and the TSC1/TSC2 tumor suppressor complex. *Genes Dev.* **18**, 2893–2904 (2004).
29. Liu, L. *et al.* Hypoxia-induced energy stress regulates mRNA translation and cell growth. *Mol. Cell* **21**, 521–531 (2006).
30. Dufner, A. & Thomas, G. Ribosomal S6 kinase signaling and the control of translation. *Exp. Cell Res.* **253**, 100–109 (1999).
31. Harrington, L. S., Findlay, G. M. & Lamb, R. F. Restraining PI3K: mTOR signalling goes back to the membrane. *Trends Biochem. Sci.* **30**, 35–42 (2005).
32. Martin, D. E. & Hall, M. N. The expanding TOR signaling network. *Curr. Opin. Cell Biol.* **17**, 158–166 (2005).
33. Long, X., Ortiz-Vega, S. & Avruch, J. Rheb binding to mammalian target of rapamycin (mTOR) is regulated by amino acid sufficiency. *J. Biol. Chem.* **280**, 23433–23436 (2005).
34. Kim, J. E. & Chen, J. Cytoplasmic-nuclear shuttling of FKBP12-rapamycin-associated protein is involved in rapamycin-sensitive signaling and translation initiation. *Proc. Natl Acad. Sci. USA* **97**, 14340–14345 (2000).
35. Park, I. H., Bachmann, R., Shirazi, H. & Chen, J. Regulation of ribosomal S6 kinase 2 by mammalian target of rapamycin. *J. Biol. Chem.* **277**, 31423–31429 (2002).
36. Paglin, S. *et al.* Rapamycin-sensitive pathway regulates mitochondrial membrane potential, autophagy, and survival in irradiated MCF-7 cells. *Cancer Res.* **65**, 11061–11070 (2005).
37. Jensen, K., Shiels, C. & Freemont, P. S. PML protein isoforms and the RBCC/TRIM motif. *Oncogene* **20**, 7223–7233 (2001).
38. Corada, M. *et al.* A monoclonal antibody to vascular endothelial-cadherin inhibits tumor angiogenesis without side effects on endothelial permeability. *Blood* **100**, 905–911 (2002).
39. Liao, F. *et al.* Selective targeting of angiogenic tumor vasculature by vascular

- endothelial-cadherin antibody inhibits tumor growth without affecting vascular permeability. *Cancer Res.* **62**, 2567–2575 (2002).
40. Zhu, J., Lallemand-Breitenbach, V. & de Thé, H. Pathways of retinoic acid- or arsenic trioxide-induced PML/RAR α catabolism, role of oncogene degradation in disease remission. *Oncogene* **20**, 7257–7265 (2001).
 41. Lavau, C. *et al.* The acute promyelocytic leukaemia-associated PML gene is induced by interferon. *Oncogene* **11**, 871–876 (1995).
 42. Chelbi-Alix, M. K. *et al.* Induction of the PML protein by interferons in normal and APL cells. *Leukemia* **9**, 2027–2033 (1995).
 43. Sawyers, C. L. Will mTOR inhibitors make it as cancer drugs? *Cancer Cell* **4**, 343–348 (2003).
 44. Vignot, S., Faivre, S., Aguirre, D. & Raymond, E. mTOR-targeted therapy of cancer with rapamycin derivatives. *Ann. Oncol.* **16**, 525–537 (2005).
 45. Bernardi, R. *et al.* PML regulates p53 stability by sequestering Mdm2 to the nucleolus. *Nature Cell Biol.* **6**, 665–672 (2004).
 46. Sarbassov, D. D. *et al.* Rictor, a novel binding partner of mTOR, defines a rapamycin-insensitive and raptor-independent pathway that regulates the cytoskeleton. *Curr. Biol.* **14**, 1296–1302 (2004).
 47. Grisendi, S. *et al.* Role of nucleophosmin in embryonic development and tumorigenesis. *Nature* **437**, 147–153 (2005).
- Supplementary Information** is linked to the online version of the paper at www.nature.com/nature.
- Acknowledgements** We are grateful to W. Gerald for providing the prostate cancer TMAs. We thank M. Socorro Jiao and M. Drobnjak for help with immunohistochemistry, P. Burgman and S. Carlin for help with hypoxia experiments, B. Carver for help with statistical analysis, L. DiSantis, R. Hobbs and J. Clohessy for critical reading of the manuscript, and all members of the Pandolfi laboratory for comments and discussion. This work was supported by NIH grants to P.P.P.
- Author Information** Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to P.P.P. (p-pandolfi@ski.mskcc.org).

LETTERS

The rapid formation of a large rotating disk galaxy three billion years after the Big Bang

R. Genzel^{1,2}, L. J. Tacconi¹, F. Eisenhauer¹, N. M. Förster Schreiber¹, A. Cimatti^{1,3}, E. Daddi⁴, N. Bouché¹, R. Davies¹, M. D. Lehnert¹, D. Lutz¹, N. Nesvadba¹, A. Verma¹, R. Abuter¹, K. Shapiro⁵, A. Sternberg⁶, A. Renzini⁷, X. Kong⁸, N. Arimoto⁹ & M. Mignoli¹⁰

Observations and theoretical simulations have established a framework for galaxy formation and evolution in the young Universe^{1–3}. Galaxies formed as baryonic gas cooled at the centres of collapsing dark-matter haloes; mergers of haloes and galaxies then led to the hierarchical build-up of galaxy mass. It remains unclear, however, over what timescales galaxies were assembled and when and how bulges and disks—the primary components of present-day galaxies—were formed. It is also puzzling that the most massive galaxies were more abundant and were forming stars more rapidly at early epochs than expected from models^{4–7}. Here we report high-angular-resolution observations of a representative luminous star-forming galaxy when the Universe was only 20% of its current age. A large and massive rotating protodisk is channelling gas towards a growing central stellar bulge hosting an accreting massive black hole. The high surface densities of gas, the high rate of star formation and the moderately young stellar ages suggest rapid assembly, fragmentation and conversion to stars of an initially very gas-rich protodisk, with no obvious evidence for a major merger.

Imaging spectroscopy of high-redshift galaxies at high angular resolution of well understood rest-frame optical spectral diagnostics is now becoming feasible with advanced instruments on large ground-based telescopes. This promises new empirical information about the crucial epoch of galaxy evolution near a cosmological redshift $z \sim 2$, about 3 billion years after the Big Bang, when the Universe was about 20% of its present age. We have recently begun a study of a representative sample of $z \sim 2$ –3 star-forming galaxies, selected on the basis of their rest-frame ultraviolet/optical fluxes and colours, with the near-infrared integral field spectrometer SINFONI on the Very Large Telescope of the European Southern Observatory^{8,9}. Our first results¹⁰ revealed that fairly large and massive protodisk galaxies were present already at $z \sim 2$ –3. However, we did not have sufficient resolution to distinguish unambiguously between a merger and a disk interpretation, or to resolve the bulge and disk components. For one of these luminous star-forming galaxies, BzK-15504 ($z = 2.38$; refs 11, 12), the presence of a nearby star and excellent atmospheric conditions allowed us to take full advantage of the adaptive optics mode of SINFONI. We achieved an angular resolution of $\sim 0.15''$ (1.2 kpc or 4,000 light years), more than three times better than in our previous work. BzK-15504 is fairly typical of rest-frame optically bright, actively star-forming galaxies at that redshift (for details see legend to Fig. 1, and Supplementary Information). The SINFONI spectral data reveal the spatial distribution

and kinematics of spectroscopic tracers in unprecedented detail. H α line emission tracing ionized gas in many compact star-formation complexes is distributed over an extended region of diameter $\sim 1.6''$ (13 kpc), approximately centred on the continuum peak (Fig. 1). The H α surface brightness distribution is somewhat asymmetric (the emission towards the northwest is significantly brighter) but low-level H α emission is detected over a comparable area in both northwest and southeast quadrants. Similarly asymmetric distributions of the bright ionized gas are often seen in otherwise symmetric, local-Universe disk galaxies. There they reflect the instantaneous distribution of the youngest star-forming regions and, in some cases, of dust extinction^{13,14}. The velocity field of H α is remarkably symmetric and is described very well by a combination of circular motion (rotation) in the outer parts and strong radial flow in the nuclear region. BzK-15504 also has a bright near-nuclear concentration of H α emission, accompanied by wider lines of 450 km s^{-1} full-width at half-maximum (FWHM) (Figs 1 and 2, and Supplementary Information).

In the outer regions ($r > 0.4''$) the highest-velocity and lowest-velocity H II region complexes are found at position angle $\sim 24^\circ$ west of north, which is also the morphological major axis. Along this axis the absolute values of the mean line velocities (relative to the systemic velocity) increase sharply with distance from the centre and reach a constant value at $|r| > 0.4''$ (Fig. 2b). Along the minor axis in the blue-shifted northwestern quadrant the absolute values of the mean velocities decrease smoothly on each side of the major axis (Figs 1 b–i and 2d). Taken together, these characteristics constitute compelling qualitative evidence for rotation in a disk inclined with respect to the sky plane¹⁵. For quantitative modelling we adopted an azimuthally symmetric exponential disk (see Supplementary Information), which fits the data very well. The best-fitting models from χ^2 minimization (Table 1) have a radial exponential scale length of 4.5 kpc, a maximum circular velocity of 230 km s^{-1} and an overall dynamical mass of 1.1×10^{11} solar masses ($M_\odot$) within $r \sim 8 \text{ kpc}$. Although the ionized gas in BzK-15504 is therefore clearly a large rotating disk, the question arises whether this conclusion also holds for the underlying galaxy as a whole, or whether a merger configuration might not also be possible. The possibility of a major merger (a mass ratio of the two galaxies of less than 3:1) is not compatible with the observed kinematics. For two galaxy mass centres separated on the same scale as the gas disk (4 kpc or more), the gas dynamics would be strongly perturbed and would deviate substantially from that of a simple disk. For a very advanced major merger, most of the

¹Max-Planck-Institut für extraterrestrische Physik (MPE), Giessenbachstr. 1, 85748 Garching, Germany. ²Department of Physics, Le Conte Hall, University of California, 94720 Berkeley, California, USA. ³Istituto Nazionale di Astrofisica, Osservatorio Astrofisico di Arcetri, Largo E. Fermi 5, I-50125 Firenze, Italy. ⁴National Optical Astronomy Observatory, 950 North Cherry Avenue, Tucson, Arizona 85719, USA. ⁵Department of Astronomy, Campbell Hall, University of California, Berkeley 94720, California, USA. ⁶School of Physics and Astronomy, Tel Aviv University, Tel Aviv 69978, Israel. ⁷Osservatorio Astronomico di Padova, Vicolo dell'Osservatorio 5, I-35122 Padova, Italy. ⁸Center for Astrophysics, University of Science and Technology of China, 230026 Hefei, PR China. ⁹National Astronomical Observatory, Mitaka, Tokyo 181 8588, Japan. ¹⁰Istituto Nazionale di Astrofisica, Osservatorio Astronomico di Bologna, Via Gobetti 101, I-40129 Bologna, Italy.

gas would have a small angular momentum and would be concentrated on the spatial scale of the (stellar) merger remnant (less than 1 kpc) even if a small fraction of loosely bound gas were in a rotating disk with large angular momentum. Finally, in a minor merger (a mass ratio of more than 3:1) the more massive galaxy, including its gas distribution, can retain a significant angular momentum and resemble a rotating disk. This case is compatible with our observations. We conclude that BzK-15504 is a large and massive protodisk galaxy with an angular momentum comparable to that of local Universe Sb/c spirals.

Several properties of the source indicate that the protodisk was assembled rapidly, on a timescale of a few hundred million years, and has been converting a significant fraction of its total baryonic mass to stars on that timescale, with no obvious evidence for a major merger. The dynamical properties and the global and local matter surface densities (Table 1) show that the gas in the protodisk is unstable to global star formation ('Toomre' Q parameter ≤ 1 ; refs 16, 17) and fragmentation¹⁸. If the local gas motions seen towards the star-forming complexes in Fig. 1 are virialized (see the legend to Fig. 3, and Supplementary Information) the characteristic diameters (~ 0.1 – $0.2''$) and intrinsic local velocity dispersions of ~ 30 – 60 km s^{-1} imply very large complex masses of $(2\text{--}10) \times 10^8 M_\odot$, which in turn are exactly what would be expected for a globally unstable disk. From the $\text{H}\alpha$ line flux and a 'Kennicutt'¹⁹ conversion from $\text{H}\alpha$ luminosity to star formation rate and modelling of the ultraviolet spectral energy distribution (Supplementary Information) we infer a total star-formation rate of $100\text{--}200 M_\odot \text{ yr}^{-1}$

for a Chabrier–Kroupa^{20,21} stellar mass function (Table 1). This is close to the 'maximum' rate²² for forming the current stellar mass of $\sim 8 \times 10^{10} M_\odot$ (Table 1) on a timescale of $\sim 500 \text{ Myr}$, the stellar age derived from a fit to the spectral energy distribution of the object (see Supplementary Information). Further, the ratio of circular velocity to local gas velocity dispersion in BzK-15504 is $v_c/\sigma \sim 2\text{--}4$ (all corrected for instrumental effects and inclination), as in three other such galaxies in ref. 10 (see Supplementary Information). The thin disks of local spiral galaxies have much greater ratios: $v_c/\sigma \sim 10\text{--}50$. The ionized gas disk is dynamically 'hot'. The energy source for this large dispersion is plausibly due to 'feedback' through supernova explosions, or to radiation and winds from the active star formation itself, or to the tapping of the accretion energy in the disk's assembly^{10,23–25}. All these scenarios require a fairly rapid (a few hundred million years) formation timescale, consistent with the arguments given above. As discussed above, the remarkable feature of BzK-15504 is that the symmetric properties of the observed outer velocity field argue against a recent major merger event of two galaxies as the trigger of this intense star-formation activity. Our data imply a more quiescent but still fast gas inflow from the dark halo, or a rapid series of minor mergers. Combining the gas and stellar masses from Table 1 accounts fully for the inferred dynamical mass within $\sim 10 \text{ kpc}$. The disk seems to be dominated by baryons.

Although the outer disk is well described by circular motion, velocity field and data-model difference maps of the inner disk show strong deviations from rotation within $\sim 0.4''$ (3 kpc) of the nucleus. This conclusion is apparent on inspection of the residual maps in

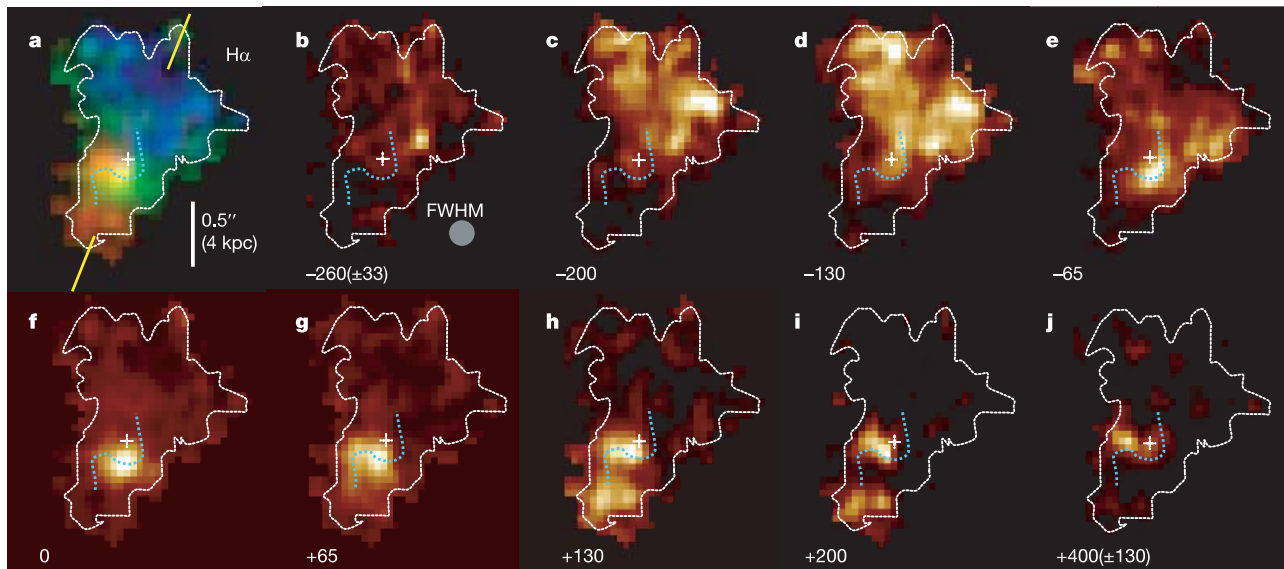


Figure 1 | Velocity maps of $\text{H}\alpha$ emission in BzK-15504 ($z = 2.3834$). **a**, An RGB composite colour map of the entire $\text{H}\alpha$ emission, with three colours encoding the total velocity range of the emission. **b–i**, Velocity maps summed over 65 km s^{-1} . **j**, The highly red-shifted feature east of the nucleus (260 km s^{-1} range). The velocity maps **b–e** show that the gas with the largest blueshift is found near the major axis, but as the velocity becomes less blue the emission moves farther and farther away on each side of this axis. This pattern is a clear characteristic of a rotating disk. The crosses in the panels mark the position of the continuum peak. The dotted thin white curve outlines the shape of the integrated $\text{H}\alpha$ emission at the $\sim 20\%$ level. The dotted light blue S-shape indicates the radial flow discussed in the text. The two yellow lines mark the orientation of the major axis of the disk. North is up and east is to the left, and a bar marks a length of $0.5''$. The source BzK-15504 was discovered as part of a K-band ($1.9\text{--}2.4 \mu\text{m}$, rest frame $\sim 6,500 \text{ \AA}$) survey (to $K_{\text{AB}} < 21.9$) in the RA = 11 h, dec. = -21° 'deep 3a-F' Subaru field¹² (Supplementary Information). Star-forming galaxies at $z \sim 2$ were selected from the so-called '(s)BzK' colour criteria (Supplementary Information). Follow-up spectroscopy was then performed with VIMOS on

the ESO VLT. The source BzK-15504 ($K_{\text{AB}} = 21.08$, $z = 2.38$) is among the brightest 30% of the $z \geq 2$ BzK galaxies. The rest-frame ultraviolet spectrum of the source (Supplementary Fig. 2a) exhibits strong Ly α , C IV $1,549\text{-\AA}$ and C III] $1,909\text{-\AA}$ emission, indicative of the presence of a central AGN. The SINFONI data were taken in two sets of 6-h integrations in the K-band, both using a nearby $V = 16.3$ star as reference for the MACAO adaptive optics system. SINFONI delivers spectra simultaneously over a contiguous two-dimensional field of $64 \text{ pixels} \times 32 \text{ pixels}$. The FWHM spectral resolution is 54 km s^{-1} . The data set shown in Figs 1–3 was taken in excellent atmospheric seeing (FWHM on the visible guider $\sim 0.4\text{--}0.6''$) and coherence time ($7\text{--}10 \text{ ms}$) conditions and used the $0.05'' \times 0.1''$ pixel scale. The resulting FWHM angular resolution is $0.15''$, corresponding to 1.2 kpc ($4,000$ light years). Another data set was taken in the $0.125'' \times 0.25''$ pixel scale and resulted in a FWHM spatial resolution of $0.45''$. The SINFONI data were reduced and calibrated with the custom-developed 'SPRED'^{18,29} software, with additional tools to treat very faint signals and analyse the final data cubes. All physical units in this paper are based on a concordance, flat Λ CDM (lambda-cold dark matter) cosmology with $H_0 = 70 \text{ km s}^{-1} \text{ Mpc}^{-1}$.

Table 1 | Physical properties of BzK-15504

Parameter	Value	Description
Z	2.3834	Look-back time 10.7 Gyr, $1'' = 8.135$ kpc
K_s	19.2	Total K_s band magnitude (Vega system, uncorrected for 30% line emission)
A_V	$0.9(+0.3,-0.3)$	Extinction derived from spectral energy distribution fitting (SED)
R_*	$140(+110,-80)M_\odot\text{yr}^{-1}$	Star-formation rate from H α flux of $2.5 \times 10^{-16}\text{erg s}^{-1}\text{cm}^{-2}$ (ref. 19) and ultraviolet SED, for IMF in refs 20, 21 and $A_V = 0.9$ mag
t_*	$5(+5,-2) \times 10^8\text{yr}$	Stellar age from SED fitting, for continuous or $3 \times 10^8\text{-yr}$ duration burst
Σ_*	$1.2M_\odot\text{yr}^{-1}\text{kpc}^{-2}$	Star-formation rate surface density
M_{dyn}	$(11.3 \pm 1) \times 10^{10}M_\odot$	Dynamical mass within $r = 1.1''$, corrected for inclination $i = 48 \pm 3^\circ$
M_*	$7.7(+3.9,-1.3) \times 10^{10}M_\odot$	From SED and IMF from refs 20, 21, excluding stellar mass loss
M_{gas}	$4.3 \times 10^{10}M_\odot$	Total gas mass from H α and Schmidt-Kennicutt law ¹⁹ , for $A_V = 0.9$ mag
Σ_{gas}	$350M_\odot\text{pc}^{-2}$	Total gas surface density from H α surface brightness and Schmidt-Kennicutt law ¹⁹
V_c	$230 \pm 16\text{km s}^{-1}$	Circular velocity at $r = 5\text{-}10$ kpc
$R_{1/e}$	$4.5 \pm 1\text{kpc}$	Radial scale length of H α disk
$Z_{1/e}$	$1 \pm 0.5\text{kpc}$	Vertical scale length of H α disk, from $v_c/\sigma = 3 \pm 1$
Q	0.8 ± 0.4	Toomre Q parameter for global disk stability ¹⁶

See Supplementary Information for details; quoted uncertainties are 1 standard deviation of the fit.

Fig. 3, which exhibit significant velocity residuals southwest of the nucleus. The origin of these residuals is gas at slightly blue-shifted velocities entering the nuclear region from the northwest, then passing along the minor axis below the nuclear position and exiting southeast into the red-shifted part of the disk (see Fig. 1a, d–g, shown by the dotted S-shaped curve). This trend is reflected in the overall velocity field of Fig. 3a as a twist of the iso-velocity contours, reaching an amplitude of 70–120 km s^{−1} in the velocity residual map (Fig. 3c). Very similar patterns are seen in many local-Universe galaxies (see, for example, refs 15, 26), where they are interpreted as the tell-tale signature of radial inward streaming of gas from the disk into the nucleus (for example, in response to a stellar bar). This inflow could be an important contributor to the growth of a substantial bulge whose presence is already apparent as a bright stellar peak near the centre of BzK-15504 (see Supplementary Information). In addition

there is a red-shifted high-velocity tail (to ~500 km s^{−1}) of the H α emission ~0.2'' east of the nuclear continuum peak (Fig. 1j). This feature is also prominent in [N II] (Supplementary Information) and may be powered by hard ultraviolet radiation from an active galactic nucleus (AGN), whose presence in BzK-15504 at that position is indicated by the properties of its ultraviolet spectrum (Supplementary Information). Inward nuclear flow and a contribution from an AGN plausibly explain the relatively large nuclear velocity dispersion (Fig. 2c) that is not accounted for by the disk rotation model.

Our observations indicate a self-consistent picture of rapid gas

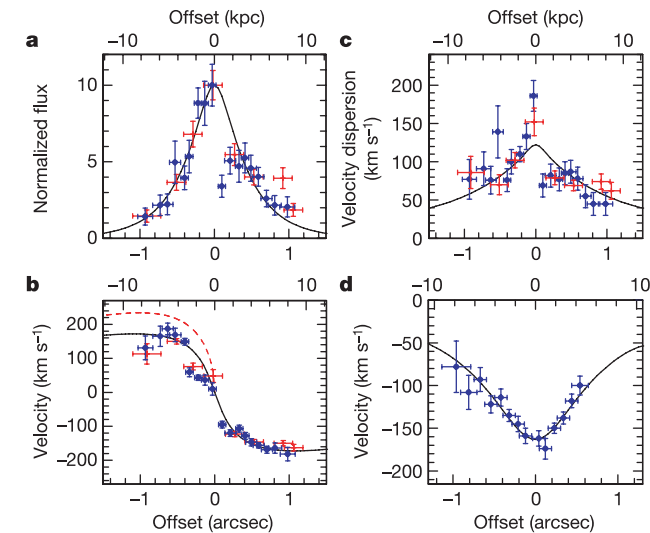


Figure 2 | Velocity and intensity distributions along the major and minor axes of the source. **a**, H α intensity along the morphological and kinematic major axis at position angle 24° west of north. **b**, **c**, Peak velocity (**b**) and velocity dispersion (**c**) extracted along the major axis by fitting gaussians to spectra in selected apertures. **d**, Peak velocity in selected apertures along the minor axis (position angle 114° west of north) through a position 0.69'' northwest of the centre and averaged over 0.25'' along the major axis. Filled blue circles denote spectra in the 0.15'' high resolution data of Fig. 1; red crosses represent the second independent data set at 0.45'' spatial resolution. Continuous curves denote the best fitting exponential disk model to the data. The dotted curve in **b** is the inclination and resolution corrected, intrinsic rotation curve inferred from our modelling. In all panels vertical error bars represent formal 1- σ uncertainties in the measurements. Horizontal error bars indicate the diameter of the synthetic apertures.

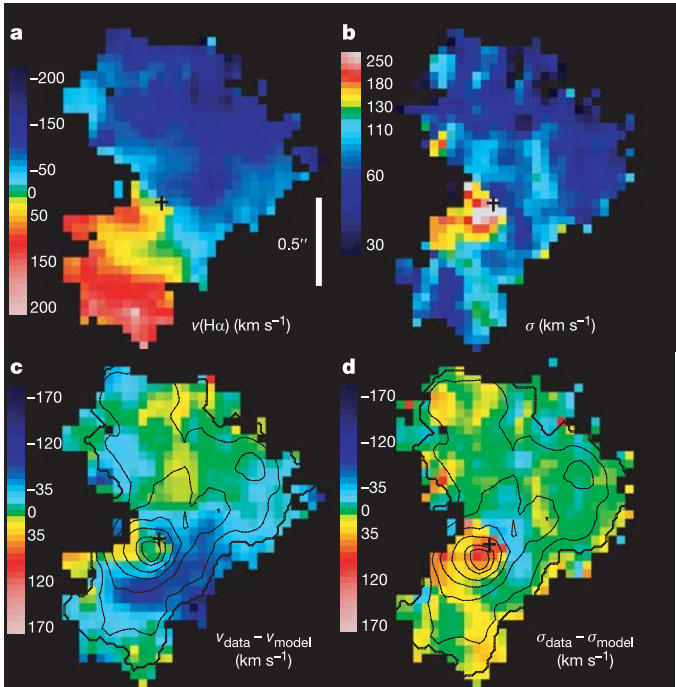


Figure 3 | Two-dimensional distributions of first and second moments of the H α velocity distribution. **a**, Extracted mean velocity map. **b**, Extracted velocity dispersion map. **c**, **d**, Difference maps of **a** and **b**, respectively, and the corresponding best-fitting exponential disk model distributions. Superposed are contours of integrated H α emission. In all cases the data and models were smoothed to 0.19'' FWHM. The crosses denote the position of the continuum peak. The strong deviations (in **a** and **c**) near the dynamical centre of the velocity field from that of the simple rotation pattern in the outer disk indicate a 70–120-km s^{−1} component of radial motion, either inflow or outflow. The spatial connection of this radially streaming gas to the outer disk apparent from the channel maps in Fig. 1 strongly favours radial inflow.

inflow from the halo, followed by gravitational instability and star formation in a massive, gas-rich protodisk, and a subsequent build-up of a central bulge through inflow triggered by disk instabilities and/or minor mergers^{18,27}. Although BzK-15504 is the only $z \sim 2$ galaxy for which such high-resolution data are currently available, its global properties, including its star-formation rate, stellar and dynamical masses and size, are similar to the general population of actively star-forming galaxies at $z \sim 2$ (Supplementary Information). Thus, the phenomena observed in this galaxy, such as rapid gas inflow, gravitational instability and subsequent build-up of a central bulge, are probably of general importance for the rapid formation and assembly of massive (comparable to, or more massive than, the Milky Way) galaxies in this critical cosmic epoch^{6,7}. If the large thickness of the H α gas layer inferred from the observed $v/\sigma \sim 3$ ratio is characteristic of the underlying stellar disk, an obvious question raised by our study is whether we are witnessing the formation of a 'thick' disk, possibly related to the so-called 'old thick' stellar disks seen in several spiral galaxies at low redshift²⁸. Alternatively, the $z \sim 2$ protodisks might be destroyed later by major mergers and contribute to the formation of elliptical/S0 galaxies.

Received 25 April; accepted 6 July 2006.

- White, S. D. M. & Rees, M. J. Core condensation in heavy halos—a two-stage theory for galaxy formation and clustering. *Mon. Not. R. Astron. Soc.* **183**, 341–358 (1978).
- Davis, M., Efstathiou, G., Frenk, C. S. & White, S. D. M. The evolution of large-scale structure in a universe dominated by cold dark matter. *Astrophys. J.* **292**, 371–394 (1985).
- Ellis, R. in *Galaxies at High Redshift* (eds Perez-Fournon, I., Balcells, M., Moreno-Insertis, F. & Sanchez, F.) 1–28 (Cambridge Univ. Press, Cambridge, 2003).
- Fontana, A. et al. The K20 survey. VI. The distribution of the stellar masses in galaxies up to $z \geq 2$. *Astron. Astrophys.* **424**, 23–42 (2004).
- Chapman, S. C., Blain, A. W., Smail, I. & Ivison, R. J. A redshift survey of the submillimeter galaxy population. *Astrophys. J.* **622**, 772–796 (2005).
- Cimatti, A. et al. Old galaxies in the young Universe. *Nature* **430**, 184–187 (2004).
- Glazebrook, K. et al. A high abundance of massive galaxies 3–6 billion years after the Big Bang. *Nature* **430**, 181–184 (2004).
- Eisenhauer, F. et al. SINFONI—Integral field spectroscopy at 50 milli-arcsecond resolution with the ESO VLT. *SPiE* **4841**, 1548–1561 (2003).
- Bonnet, H. et al. First light of SINFONI at the VLT. *Messenger* **117**, 17–24 (2004).
- Förster Schreiber, N. M. et al. SINFONI integral field spectroscopy of $z \sim 2$ UV-selected galaxies: rotation curves and dynamical evolution. *Astrophys. J.* **645**, 1062–1075 (2006).
- Daddi, E. et al. A new photometric technique for the joint selection of star-forming and passive galaxies at $1.4 \leq z \leq 2.5$. *Astrophys. J.* **617**, 746–764 (2004).
- Kong, X. et al. A wide area survey for high-redshift massive galaxies. I. Number counts and clustering of BzKs and EROs. *Astrophys. J.* **638**, 72–87 (2006).
- Ganda, K. et al. Late-type galaxies observed with SAURON: two-dimensional stellar and emission-line kinematics of 18 spirals. *Mon. Not. R. Astron. Soc.* **367**, 46–78 (2006).
- Daigle, O. et al. α kinematics of the SINGS nearby galaxies survey. I. *Mon. Not. R. Astron. Soc.* **367**, 469–512 (2006).
- van der Kruit, P. & Allen, R. The kinematics of spiral and irregular galaxies. *Annu. Rev. Astron. Astrophys.* **16**, 103–139 (1978).
- Toomre, A. On the gravitational stability of a disk of stars. *Astrophys. J.* **139**, 1217–1238 (1964).
- Kennicutt, R. C. The star formation law in galactic disks. *Astrophys. J.* **644**, 685–703 (1989).
- Immeli, A., Samland, M., Gerhard, O. & Westera, P. Gas physics, disk fragmentation and bulge formation in young galaxies. *Astron. Astrophys.* **413**, 547–561 (2004).
- Kennicutt, R. C. The global Schmidt Law in star-forming galaxies. *Astrophys. J.* **498**, 541–552 (1998).
- Kroupa, P. On the variation of the initial mass function. *Mon. Not. R. Astron. Soc.* **322**, 231–246 (2001).
- Chabrier, G. Galactic stellar and substellar initial mass function. *Publ. Astron. Soc. Pacif.* **115**, 763–795 (2003).
- Tacconi, L. J. et al. High-resolution millimeter imaging of submillimeter galaxies. *Astrophys. J.* **640**, 228–240 (2006).
- Dib, S., Bell, E. & Burkert, A. The supernova rate-velocity dispersion relation in the interstellar medium. *Astrophys. J.* **638**, 797–810 (2006).
- Thompson, T. A., Quataert, E. & Murray, N. Radiation pressure-supported starburst disks and active galactic nucleus fueling. *Astrophys. J.* **630**, 167–185 (2005).
- Monaco, P. Physical regimes for feedback in galaxy formation. *Mon. Not. R. Astron. Soc.* **352**, 181–204 (2004).
- Lindblad, P.-O. NGC 1365. *Astron. Astrophys. Rev.* **9**, 221–271 (1999).
- D'Onglia, E., Burkert, A., Murante, G. & Khochfar, S. How galaxies lose their angular momentum. *Mon. Not. R. Astron. Soc.* (submitted); preprint at (<http://arXiv.org/astro-ph/0602005>) (2006).
- Yoachim, P. & Dalcanton, J. J. Structural parameters of thin and thick disks in edge-on disk galaxies. *Astron. J.* **131**, 226–249 (2006).
- Schreiber, J. et al. in *Astronomical Data Analysis Software and Systems (ADASS) XIII* (eds Ochsenbein, F., Allen, M. G. & Egret, D.) 380–383 (ASP Conf. Proc. 314, San Francisco, 2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the SINFONI team members, from the ESO and MPE, for their work, which made these observations possible; the Paranal staff, especially J. Navarrete and P. Amico for their support; and A. Burkert, O. Gerhard and P. Monaco for discussions. A.C. acknowledges support through a Bessel Prize of the Alexander von Humboldt Foundation. This study is based on observations at the Very Large Telescope (VLT) of the ESO, Paranal, Chile.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.G. (genzel@mpe.mpg.de).

LETTERS

No signature of clear CO₂ ice from the 'cryptic' regions in Mars' south seasonal polar cap

Yves Langevin¹, Sylvain Douté², Mathieu Vincendon¹, François Poulet¹, Jean-Pierre Bibring¹, Brigitte Gondet¹, Bernard Schmitt² & F. Forget³

The seasonal polar ice caps of Mars are composed mainly of CO₂ ice^{1,2}. A region of low (<30%) albedo has been observed within the south seasonal cap during early to mid-spring^{3,4}. The low temperature of this 'cryptic region' has been attributed to a clear slab of nearly pure CO₂ ice, with the low albedo resulting from absorption by the underlying surface⁴. Here we report near-infrared imaging spectroscopy of the south seasonal cap. The deep and broad CO₂ absorption bands that are expected in the near-infrared with a thick transparent slab of CO₂ ice are not observed. Models of the observed spectra indicate that the low albedo results from extensive dust contamination close to the surface of a CO₂ ice layer, which could be linked to atmospheric circulation patterns^{5,6}. The strength of the CO₂ absorption increases after mid-spring, so part of the dust is either carried away or buried more deeply in the ice layer during the CO₂ ice sublimation process.

The most comprehensive observations of the cryptic region to date⁴ were obtained with the Thermal Emission Spectrometer (TES) on board the Mars Global Surveyor (MGS) spacecraft, which was able to monitor surface temperatures down to CO₂ ice sublimation temperatures (140 K) in the thermal infrared as well as the albedo in the visible region with a broadband filter. These observations revealed that the cryptic region is a major feature of the south seasonal cap during early to mid-spring (heliocentric longitude L_s , 200°–230°), and at that time extends from 60° E to 210° E at latitudes around 80° S. The formation of a clear slab of nearly pure CO₂ ice in the cryptic region (which would explain the low temperature and albedo⁴) has been linked with weather systems originating in the southern basins, Hellas and Argyre⁵.

The south seasonal cap was observed in winter and spring by the OMEGA visible/near-infrared imaging spectrometer on board Mars Express from December 2004 to May 2005 at slant distances ranging from 2,000 km to 8,000 km, so that each OMEGA observation resulted in a 128-pixel-wide track with a pixel size of 2.4 km to 10 km. Major absorption bands of both CO₂ and H₂O ice^{7–9} can be observed in the spectral range of OMEGA (0.35 μ m to 5.1 μ m wavelength). The mean path of photons in icy material can be derived from the relative band depths, through laboratory simulations and modelling. The average grain size of the CO₂ ice, as well as the amount of dust and H₂O-ice inclusions, can then be evaluated¹⁰. The strongest CO₂ ice signatures observed by OMEGA (Fig. 1) were obtained at latitudes of ~60° S in mid-winter (L_s ~140°). The low reflectance factor of these areas then results from absorption by the underlying surface below a clear slab of nearly pure CO₂ ice, similarly to what was proposed for the cryptic region in early spring⁴.

After the southern spring equinox (L_s , 180°), the seasonal cap could be observed all the way to the south pole, and OMEGA monitored the

evolution of the cryptic region from the equinox to the end of CO₂ ice sublimation. An extensive coverage was obtained with ten OMEGA swaths in early June 2005, from L_s 221.6° to 224.4° (Fig. 2). The low albedo region which is observed by OMEGA from 60° E to 210° E around 80° S is remarkably consistent with previous observations of the cryptic region⁴ in the visible at similar values of L_s . The seasonal cap exhibits a wide range of spectral signatures (Fig. 3). A major finding is that the cryptic region is characterized by very weak signatures of CO₂ ice, which are not consistent with a thick transparent slab of CO₂ ice. This cannot result from a major fraction of sub-pixel ice-free areas, as their radiative equilibrium temperature would be ≥ 230 K, in contradiction to TES observations (<150 K)⁴ and the upper limit derived from OMEGA from emissivity at 5 μ m

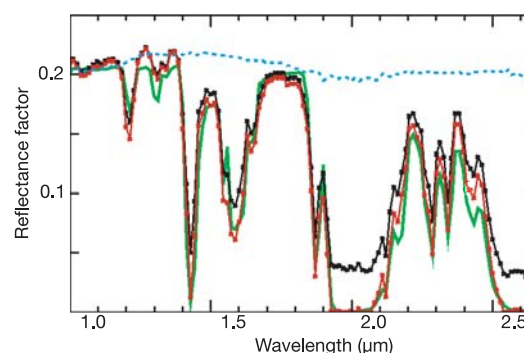


Figure 1 | Reflectance spectrum of a region at 344° E, 59° S observed by OMEGA at L_s 142°. Major CO₂ bands (1.8–2.2 μ m, >2.6 μ m) are saturated in the observed spectrum (black curve), with a significant contribution from aerosol scattering as the solar incidence is high (75°). The contribution of aerosols can be estimated from best-fit models of the observed reflectance factor at 2.65 μ m within the range of grain size and scattering properties constrained by ground-based and orbital observations^{16–18} (see Supplementary Information). The red curve is the reflectance spectrum corrected for aerosol extinction and backscattering. This spectrum has been modelled with a multilayer bidirectional reflectance model^{19,20} (see Supplementary Information) using CO₂ ice optical constants^{8,10}. The best fit (green curve) corresponds to a 30-cm layer of clear CO₂ ice with an upper limit of 0.02 vol.% for the dust and H₂O ice inclusion content assuming a size range of 10–100 μ m (ref. 10). Such a thickness of CO₂ ice is consistent with that inferred from neutron spectrometry²¹ and laser altimetry²². Blue dashed curve, reflectance spectrum of the same region at L_s 295°, when the CO₂ ice layer has sublimated. The low reflectance factor in the continuum at L_s 142° mainly results from photon absorption by the surface underlying the clear slab of CO₂ ice. The solar photons that are reflected at the surface of the slab are not observed with nadir pointing, which slightly reduces the observed reflectance factor. The signal to noise of OMEGA is always larger than 200 in this wavelength range.

¹Institut d'Astrophysique Spatiale, CNRS/Université Paris XI, Orsay Campus, 91405, France. ²Laboratoire de Planétologie de Grenoble, CNRS/Université Joseph-Fourier, Grenoble, 38400, France. ³Laboratoire de Météorologie Dynamique, CNRS/Université Paris VI, Paris, 75005, France.

(<200 K). An optically thick aerosol layer is not consistent with the low temperature or with Mars Orbiter Laser Altimeter (MOLA) observations¹¹.

We have modelled a representative spectrum observed in the cryptic region (Fig. 3, black curve) with a surface covered by two layers of CO₂ ice, each with different grain sizes and amounts of dust and H₂O ice inclusions (see Supplementary Information). As little as 1 mm of clear CO₂ ice at the surface would absorb most photons at 2.65 μ m. The best fit (Fig. 4b) is obtained with very high dust contamination in a thin granular upper layer above a nearly pure layer of CO₂ ice, so that most photons are scattered back by superficial dust while a few penetrate deep into CO₂ ice. The areal fraction occupied by dust in the upper layer (65% to 70%) and its albedo are the most strongly constrained parameters of the model. A lower limit of $\sim$ 100 mm can be derived for the thickness of the underlying clean CO₂ slab. The size of dust particles in the upper layer and the albedo of the underlying surface are only weakly constrained.

The cryptic region exhibits drastic changes in spectral reflectance in spring. Thirty days after equinox (L_s 197.3°, Fig. 4a), the reflectance factor at 1.08 μ m (33%) is already lower than that of most of the seasonal cap (albedo >40%) and CO₂ ice signatures are also weaker than that of bright regions of the cap. This requires significant surface contamination by dust. The maximum albedo contrast and weakest CO₂ signatures are observed at L_s 223.2° (Fig. 4b). At L_s

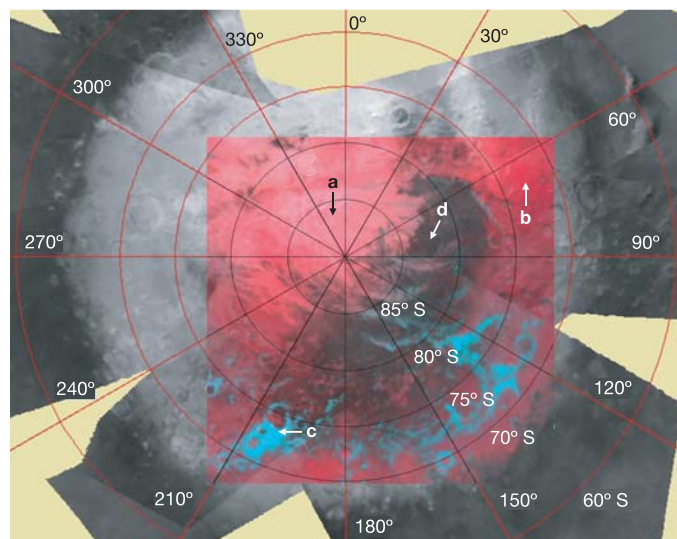


Figure 2 | Albedo of the southern seasonal cap in mid-spring at 1.08 μ m (continuum). This map is a mosaic from ten OMEGA observations between L_s 221.6° and 224.4°. The albedo is evaluated assuming a perfectly diffusive surface (Lambert albedo). The grey scale extends from 10% (black) to 80% (white). The seasonal cap appears similar to that observed by MGS at L_s 220° in 1999^{3,4,15}. Over part of the image, a false colour scheme is implemented, with each of the three colour channels remaining proportional to the reflectance factor at 1.08 μ m; the red channel is enhanced proportionally to the CO₂ band depth at 1.43 μ m; and the blue and green channels are enhanced proportionally to the H₂O ice band depth at 1.5 μ m. The low albedo region within the cap very closely matches the cryptic region observed by TES⁴. Most of the seasonal cap is spectrally dominated by CO₂ ice (reddish tones), with very bright regions ('a': 345° E, 86.2° S, albedo $\sim$ 80%) and less bright regions with much stronger CO₂ ice signatures ('b': 68° E, 73° S, albedo $\sim$ 42%). The seasonal cap is compositionally very heterogeneous between longitudes 60° E and 220° E. Relatively bright streaks (albedo >40%) are spectrally dominated by either CO₂ ice (reddish tones) or H₂O ice (bluish tones; for example, region 'c' at 200° E, 72.8° S). Within the cryptic region itself, defined by albedos <30% at latitudes higher than 70° S, signatures of both CO₂ ice and H₂O ice are much weaker than those observed in other areas of the seasonal cap. A representative area (80° E, 82.5° S), with an albedo of $\sim$ 25%, is labelled 'd'.

241.9° (Fig. 4c), the albedo and CO₂ ice band depths are much larger than at L_s 223.2° (and even larger than at L_s 197°), with a significant contribution from sub-pixel ice-free areas. At L_s 252°, CO₂ ice has sublimated. The sequence of events in the cryptic region as observed by OMEGA is therefore as follows: (1) one month after the southern spring equinox, in the cryptic region, the surface layers of the seasonal CO₂ ice slab become contaminated by dust. A contribution from precipitation of atmospheric dust is plausible, as this period corresponds to storm season, and circulation models predict large inflows southeast of Hellas⁵, where the cryptic region lies. Another contribution could be provided by venting of sub-ice bubbles formed by CO₂ ice sublimation^{12,13}. This contamination reaches a peak at L_s 220° to 225°. (2) After L_s 230°, dust contamination in the top few millimetres decreases. One possibility is that CO₂ sublimation carries away the smallest dust particles, while large dust particles become embedded in CO₂ ice¹⁴. (3) During the last stages of CO₂ ice sublimation in the cryptic region (L_s 235°–260°), patches of ice-free areas grow until the whole area is ice-free. The rapid recession of the seasonal cap in this area from L_s 230° to 260° is consistent with the increased fraction of solar energy in surface layers due to dust contamination. This leads to the well known asymmetry of the south seasonal cap in late spring and early summer^{3,4,15}.

The OMEGA results set important constraints for the models proposed for the formation of martian surface features such as dark spots, 'spiders' and 'fans', which are observed at high southern latitudes¹². Recent observations in the visible and thermal infrared provide new insights on these processes¹³. In these models, dark spots, spiders and fans are formed by the venting of sub-ice gas bubbles, which result from solar heating of the surface through a transparent CO₂ ice layer. This process can contribute to contamination of the surface of CO₂ ice with dust¹³. However, the heavily dust-contaminated upper layer, which is required above a coarse-grained CO₂ ice layer for interpreting near-infrared spectra in the cryptic region at L_s 223°, reduces by a factor of at least 3 the solar energy reaching the surface compared to clear CO₂ ice. This should hamper the formation of sub-ice gas bubbles at that stage. Spots, spiders and fans are only observed over part of the cryptic region, but the link between the geographic distribution of these features and that of the CO₂ ice signatures (and hence dust contamination of

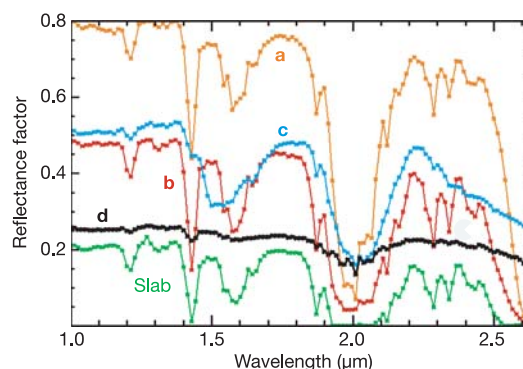


Figure 3 | Reflectance spectra of representative regions within the southern seasonal cap. The representative regions are those identified in Fig. 2 (orange, region a; red, region b; blue, region c; black, region d). Spectra are corrected for aerosol contributions and atmospheric absorption. CO₂ ice absorption bands are much deeper in region b (mean path length $\sim$ 10 cm) than in region a (mean path length $\sim$ 10 mm), hence region b is covered by much coarser CO₂ ice than region a. Strong water ice absorptions dominate the reflectance spectrum of region c. The weak CO₂ absorption bands in spectra from the cryptic region (black curve, region d) cannot correspond to a surface covered by a slab of nearly pure CO₂ ice, such as that observed by OMEGA in mid-winter, which is provided for reference (green spectrum, from Fig. 1).

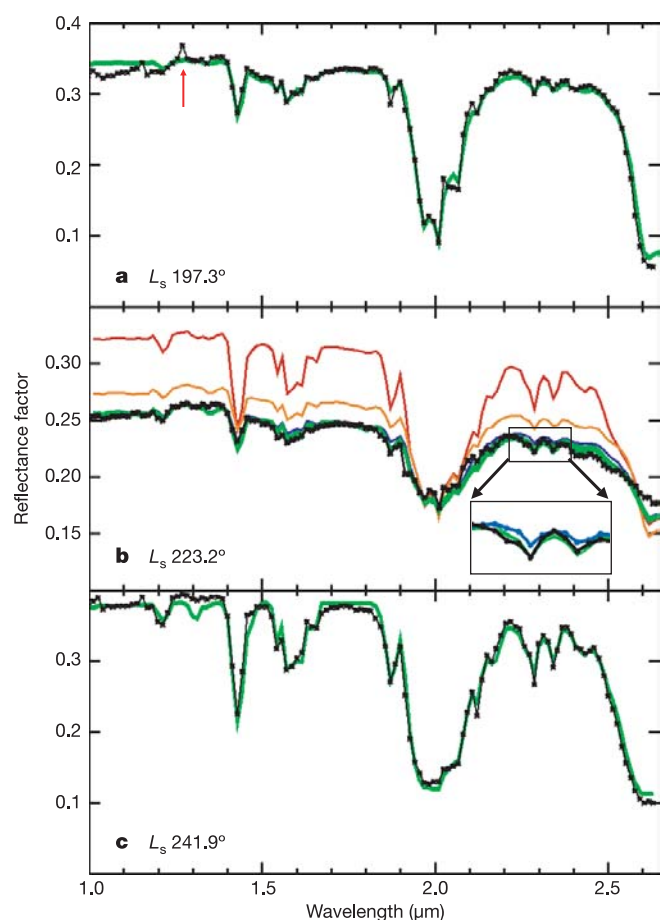


Figure 4 | Spectral evolution of a representative area in the cryptic region at 80° E, 82.5° S. **a–c,** Black lines correspond to the observed spectra at L_s 197.3° (**a**), L_s 223.2° (**b**) and L_s 241.9° (**c**). At L_s 223.2° (maximum contrast of the cryptic region), the best model fits (overlapping green curves) correspond to a thin granular upper layer 0.5–1 mm thick, with 7 wt% of 10- μ m dust particles, 0.4 wt% of 100- μ m H_2O ice particles and protruding CO_2 grains 5 mm in size, overlaying a slab 150 mm thick or more of nearly pure CO_2 ice. With a slab thickness of 75 mm (blue curve), minor CO_2 ice absorption features (inset) are too weak. If the areal fraction occupied by dust in the upper layer is reduced by a factor of 2 (red curve), more photons are scattered back from within the CO_2 layer or from the underlying surface, which leads to unacceptably strong CO_2 ice absorption features. If the same total amount of surface dust is distributed in a 2-mm-thick upper layer (orange curve), major CO_2 absorption bands also deepen. The spectrum in **a** corresponds to L_s 197.3°, but similar spectral properties are observed shortly after equinox (L_s 185°). The best fit (green curve) corresponds to a weak contamination (0.7 wt% dust, 0.06 wt% H_2O ice) of a 5-mm surface layer overlying a 150-mm-thick coarse grained layer of CO_2 ice. At L_s 241.9°, CO_2 ice bands are saturated (clear CO_2 ice slab >50 mm thick), and sub-pixel mixing with 25% of ice-free areas is required to explain the observed 10% reflectance factor in saturated bands, as an upper limit of 3% on the contribution of aerosol scattering can be obtained from observations of nearby CO_2 ice covered regions.

CO_2 ice) observed by OMEGA is not straightforward. Another issue to be settled is why dark spots, spiders and fans would not form in other regions covered by CO_2 ice in winter and spring, in particular the regions close to 60° S where OMEGA does observe clear CO_2 ice slabs (Fig. 1) during extended periods (L_s 130°–155°).

Received 17 February; accepted 13 June 2006.

- Herr, K. C. & Pimentel, G. C. Infrared absorptions near three microns recorded over the polar caps of Mars. *Science* **166**, 496–499 (1969).
- Larson, H. P. & Fink, U. Identification of carbon dioxide frost on the Martian polar caps. *Astrophys. J.* **171**, L91–L95 (1972).
- Benson, J. L. & James, P. B. Yearly comparisons of the martian polar caps: 1999–2003 Mars Orbiter Camera observations. *Icarus* **174**, 513–523 (2005).
- Kieffer, H. H. et al. Mars south polar spring and summer behaviour observed by TES: Seasonal cap evolution controlled by frost grain size. *J. Geophys. Res.* **105**, 9653–9700 (2000).
- Colaprete, A. et al. Albedo of the south pole on Mars determined by topographic forcing of atmosphere dynamics. *Nature* **435**, 184–188 (2005).
- Forget, F., Hourdin, F. & Talagrand, O. CO_2 snowfall on Mars: Simulation with a general circulation model. *Icarus* **131**, 302–316 (1998).
- Quirico, E. & Schmitt, B. Near-infrared spectroscopy of simple hydrocarbons and carbon oxides diluted in solid N_2 and as pure ices: implications for Triton and Pluto. *Icarus* **127**, 354–378 (1997).
- Grundy, W. M. & Schmitt, B. The temperature-dependent near-infrared absorption spectrum of hexagonal H_2O ice. *J. Geophys. Res.* **103**, 25809–25822 (1998).
- Hansen, G. B. Ultraviolet to near-infrared absorption spectrum of carbon dioxide ice from 0.174 to 1.8 μ m. *J. Geophys. Res.* **110**, E11003, doi:10.1029/2005JE002531 (2005).
- Bibring, J.-P. et al. Perennial water ice identified in the south polar cap of Mars. *Nature* **428**, 627–630 (2004).
- Neumann, G. A., Smith, D. E. & Zuber, M. T. Two Mars years of clouds detected by the Mars Orbiter Laser Altimeter. *J. Geophys. Res.* **108**(E4), 5023, doi:10.1029/2002JE001849 (2003).
- Piqueux, S., Byrne, S. & Richardson, M. I. Sublimation of Mars' southern seasonal CO_2 ice cap and the formation of spiders. *J. Geophys. Res.* **108**(E8), 5084, doi:10.1029/2002JE002007 (2003).
- Kieffer, H. H., Christensen, P. R. & Titus, T. N. CO_2 jets formed by sublimation beneath translucent slab ice in Mars' seasonal south polar ice cap. *Nature* doi:10.1038/nature04945 (this issue).
- James, P. B., Kieffer, H. H. & Paige, D. A. in *Mars* (eds Kieffer, H. H., Jakovsky, B. M., Snyder, C. W. & Mathews, M. S.) 934–968 (Univ. Arizona Press, Tucson, 1992).
- Titus, T. N. & Kieffer, H. H. A comparison of the Mars south polar recession rates between 1999 and 2001. *Lunar Planet. Sci. Conf.* **33**, abstr. 2071 (2002).
- Ockert-Bell, M. E., Pollack, J. B. & Forget, F. Absorption and scattering properties of the Martian dust in the solar wavelengths. *J. Geophys. Res.* **102**, 9039–9050 (1997).
- Clancy, R. T., Wolff, M. J. & Christensen, P. R. Mars aerosol studies with the MGS TES emission phase function observations: optical depths, particle sizes, and ice cloud types versus latitude and solar longitude. *J. Geophys. Res.* **108**(E9), 5098, doi:10.1029/2003JE002058 (2003).
- Drossart, P. et al. Martian aerosol properties from the Phobos/ISM experiment. *Ann. Geophys.* **9**, 754–760 (1991).
- Douté, S. & Schmitt, B. A multilayer bidirectional reflectance model for the analysis of planetary surface hyperspectral images at visible and near-infrared wavelengths. *J. Geophys. Res.* **103**, 31367–31390 (1998).
- Eluskiewicz, J., Moncet, J.-L., Titus, T. N. & Hansen, G. B. A microphysically based approach to modelling emissivity and albedo of the martian seasonal caps. *Icarus* **174**, 524–534 (2005).
- Prettyman, T. H. et al. Spatial deconvolution of Mars Odyssey neutron spectroscopy data: Analysis of Mars Southern seasonal cap. *Lunar Planet. Sci. Conf.* **36**, abstr. 1384 (2005).
- Aharonson, O. et al. Depth, distribution, and density of CO_2 deposition on Mars. *J. Geophys. Res.* **109**, E05004, doi:10.1029/2003JE002223 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We acknowledge the support of CNES, which funded the development of OMEGA and its scientific exploitation.

Author Contributions Y.L., J.-P.B. and B.G. were responsible for data reduction and observation planning; S.D., F.P. and B.S. contributed to the spectral modelling from optical constants; M.V. performed the modelling of the aerosol contribution; and F.F. provided input on the link with general circulation models.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to Y.L. (yves.langevin@ias.u-psud.fr).

CO₂ jets formed by sublimation beneath translucent slab ice in Mars' seasonal south polar ice cap

Hugh H. Kieffer^{1,2}, Philip R. Christensen³ & Timothy N. Titus²

The martian polar caps are among the most dynamic regions on Mars, growing substantially in winter as a significant fraction of the atmosphere freezes out in the form of CO₂ ice. Unusual dark spots, fans and blotches form as the south-polar seasonal CO₂ ice cap retreats during spring and summer. Small radial channel networks are often associated with the location of spots once the ice disappears. The spots have been proposed to be simply bare, defrosted ground^{1–3}; the formation of the channels has remained uncertain. Here we report infrared and visible observations that show that the spots and fans remain at CO₂ ice temperatures well into summer, and must be granular materials that have been brought up to the surface of the ice, requiring a complex suite of processes to get them there. We propose that the seasonal ice cap forms an impermeable, translucent slab of CO₂ ice that sublimates from the base, building up high-pressure gas beneath the slab. This gas levitates the ice, which eventually ruptures, producing high-velocity CO₂ vents that erupt sand-sized grains in jets to form the spots and fans and erode the channels. These processes are unlike any observed on Earth.

The vast majority of spots and fans occur in the south polar region of Mars^{1,2}, and are common in 'cryptic' terrain identified in Mars Global Surveyor Thermal Emission Spectrometer (TES) data as having CO₂ slab ice and anomalously low albedos⁴. Spots and fans (Fig. 1) are often associated with dunes but are common on mesas and individual layers of the polar layered materials^{1–3}. Spots are remarkably repeated from year to year^{2,5}, and often correspond to the centres of radially branching 'spiders' eroded into the substrate (Fig. 2)^{6,7}. Blotches are larger than spots, hundreds of metres to tens of kilometres in size, with less distinct boundaries, and, unlike most spots, have albedo patterns that are similar whether CO₂ ice is present or not.

Previous studies have suggested that these dark features are areas of early ice defrosting and exposing dark soil^{3,5,8,9}. Mars Odyssey Thermal Emission Imaging System (THEMIS)^{10–13} infrared images of the temperature of spots and fans that show that they are within 3–5 °C of the CO₂ ice temperature (~145 K), and remain at these temperatures for more than 120 days after sunrise. These temperatures are far too cold to be bare soil, which warms to >225 K within days of CO₂ ice removal^{4,14}. Thus, the spots must be on, under, or within a layer of CO₂ ice, and an alternative to defrosting is necessary to explain their formation.

We selected a representative mesa and trough system (nicknamed 'Manhattan Island') for an intense monitoring campaign (Fig. 3). In this region some spots were present at sunrise (Fig. 3a) and increased in number significantly over the next week (Fig. 3b). Some regions remained spot-free for up to 100 days then developed a large number of spots within one week (Fig. 3c, d). Fans were rarely present when spots first form, typically forming days to weeks later. Some fans lengthened by up to 1 km in multiple events, suggesting surface transport at rates well within reasonable wind velocities.

Two major brightenings occurred at Ls 201° and 241° (Ls is the aerocentric longitude of the Sun and Ls 180° is the southern spring equinox) with a contrast reversal between 'Manhattan Island' and the surroundings (Fig. 3c, e). TES albedos increased to ~0.29 and ~0.32 in dark and bright regions respectively; temperatures in all regions remained near 150 K. Many spots and fans became indistinct, suggesting the deposition of a thin veneer of bright material. Over the next several weeks the spots and fans reappeared in their previous positions, suggesting the removal of the overlying bright material or, which is less likely, the formation of new fans in the same positions.

Blotches, unlike spots, have similar albedo patterns in winter with CO₂ ice present and in summer when the ice is gone. This suggests that the ice is translucent, allowing the underlying surface albedo to be seen in winter.

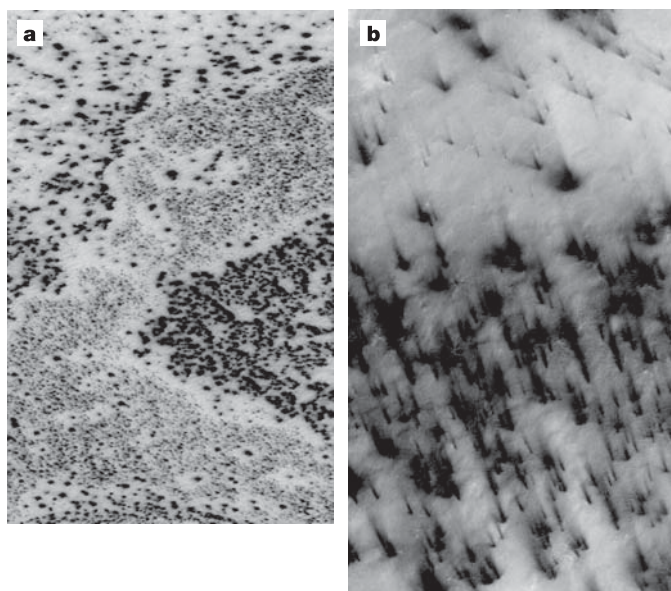


Figure 1 | Examples of spots and fans in the south polar region of Mars. **a**, MOC image E09-00231 at Ls 248° is ~3 km × 5 km in size, centred near 85.2° S and 181.4° E at ~3 m per pixel. **b**, MOC image E07-00159 at Ls 207° is ~3 km × 6 km in size centred near 86.3° S and 94.4° E at ~11 m per pixel. The dark fans are typically tens to hundreds of metres in length, 10–30° in angular size, originate from a dark spot, and point in a similar direction within a given area, sometimes with curvilinear trends influenced by topographic slopes. Occasionally fans form in multiple directions from a single spot. Some multiple fans have one diffuse edge, suggesting subsequent transport at an angle to the initial orientation. Fans are typically dark relative to the surrounding surface, although relatively bright fans also occur.

¹Celestial Reasonings, Carson City, Nevada 89703, USA. ²US Geological Survey, Flagstaff, Arizona 86001, USA. ³Department of Geological Sciences, Arizona State University, Tempe, Arizona 85287, USA.

Throughout the spring and summer the surface temperatures of spots and blotches, adjusted for the presence of a relatively warm atmosphere⁴, were only 3–5 °C warmer than CO₂ ice (~145 K) (Fig. 3). Assuming a 50/50 mixture of spots and ice within an infrared pixel, the temperature of the spots must be ≤153 K. The temperatures of the dark terrains remained near 150 K until the ice began to disappear at Ls 260° (~160 days after sunrise), at which time the surface rapidly warmed to temperatures consistent with bare soil (~235 K); see Fig. 3f. The duration of ice persistence requires an ~1-m-thick slab¹⁴, consistent with elevation, neutron and gamma-ray observations^{15–17}.

The observation that spots are only 3–5 °C warmer than CO₂ ice requires that they be on the surface in order to produce the observed warming, yet they can only be ~1 mm thick to remain in adequate conductive contact with the CO₂ ice. The occurrence of fans with their apexes originating at spots and the similarity of their orientations indicate material on the surface that is transported by wind^{6,7}; the orientations in Manhattan match the expected slope wind flow patterns towards the deep trough of Chasma Australe (Fig. 3).

The similarity in albedo patterns between winter and summer both regionally and locally (that is, blotches) suggests that in some regions the ice is transparent or translucent to a sufficient degree that surface albedo contrasts are visible through the ice. This would not be the case for a fine-grained, highly scattering CO₂ frost. There is also evidence from the infrared spectral properties that the CO₂ ice in the cryptic region is much larger-grained than the typical CO₂ frosts, and may be in the form of a translucent slab^{4,18}.

A model has been developed for the formation of spots, fans and blotches that accounts for their observed characteristics^{6,19} (Fig. 4). Dark, granular materials lie at the surface during summer. Approximately one metre of CO₂ ice^{15,16,20–22} forms the residual cap during winter (Fig. 4a). This ice anneals to form a slab^{4,6,18}. With spring sunrise, the fine dust particles incorporated during cap condensation are heated by sunlight and become mobile through the slab; these dust particles can exit the slab after ~20 martian days (sols), creating an impermeable, translucent layer that allows sunlight to penetrate to the subsurface and, in the case of blotches, allows the substrate to be visible (Fig. 4b). Insolation reaches and heats the substrate, leading to sublimation of the CO₂ ice from the base. Basal sublimation results in pressures at the base of the impermeable slab that levitate the cap. Eventually the slab ruptures locally to form jets of CO₂ gas; this gas converges towards the jets beneath the slab at lateral gas velocities of



Figure 2 | Dendritic ‘spiders’ that form beneath spots in the south polar cryptic region. This image was acquired in Mars’ southern hemisphere summer (Ls 270°) after the CO₂ ice had completely sublimated. The MOC image (R09-04023) is ~3 km × 4 km in size, centred near –86.3°S, 99.5°E at 4.3 m per pixel.

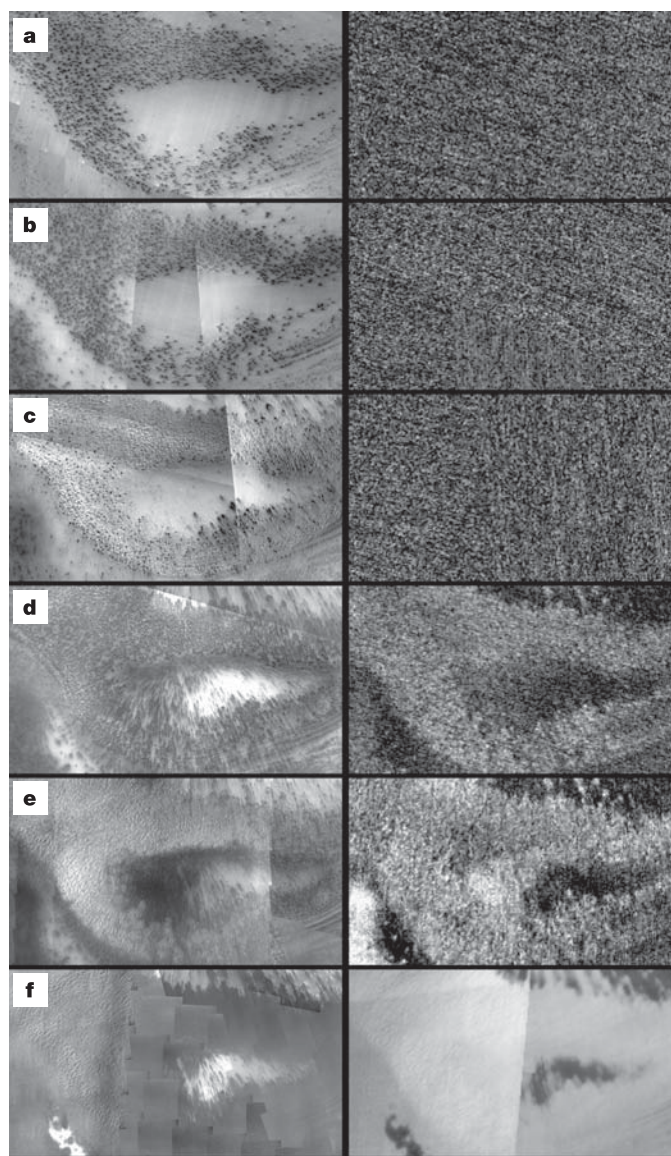


Figure 3 | The time evolution of spots and fans of a selected region of interest as seen in THEMIS visible and thermal infrared images. The ‘Manhattan Island’ region is centred at 99°E, –86.25°S. These images are a subset of over 200 THEMIS infrared and visible image pairs obtained from the end of southern winter (Ls = 176°) through midsummer (Ls 315°). This region follows the classic TES ‘cryptic’ behaviour of low albedo while remaining near the CO₂ ice temperature⁴. Simultaneous TES visible bolometer data gave well-calibrated albedo measurements at 5-km scales^{13,28}. Significantly, the spot-free ice typically had a TES albedo of 0.23–0.25; spot-covered areas were only slightly darker with an albedo of 0.22–0.24. Assuming 30% spot cover, the albedo of the spots themselves is ~0.21. The high precision of the THEMIS camera allows these subtle albedo differences to be highly exaggerated. The visible albedo (left column) in all images varies from ~0.21 (black) to 0.30 (white). The temperature (right column) varies from 145 (black) to 155 K (white) in **a** to **e**. The temperatures in **f**, after the CO₂ is mostly gone, vary from 145 K (black) to 235 K (white). Temperatures in all cases have been corrected to remove the atmospheric component. Despite the formation and evolution of dark spots and fans (**a**, Ls 182°; **b**, Ls 190°; **c**, Ls 201°) the temperatures throughout the region are within ~3 °C of the CO₂ ice temperature. Even when dark spots and fans are abundant (**d**, Ls 229°; **e**, Ls 241°), the maximum temperatures are ~150 K. Therefore, the dark materials must be in thermal contact with CO₂ ice and be on top of the cap. When the surface has completely defrosted (**f**, Ls 268°), the temperatures rapidly rise to >255 K.

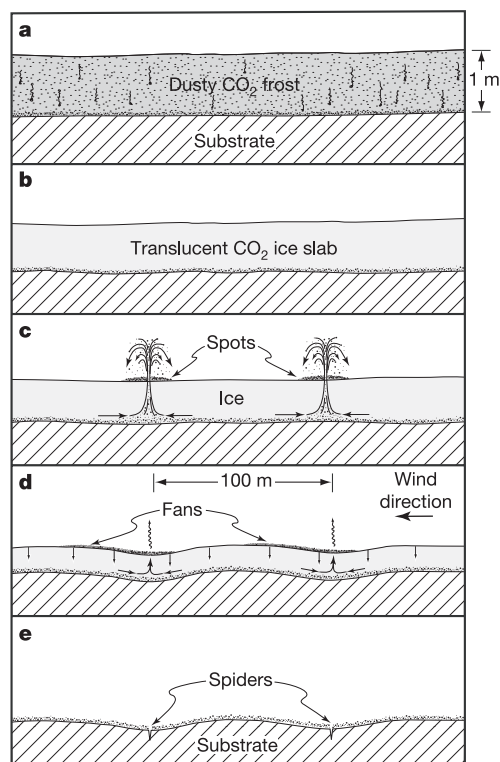


Figure 4 | A model for the formation of polar spots, fans and spiders. CO₂ ice formed during winter (a) self-cleans and anneals to form a translucent slab (b). Jets form as the pressure builds at the base of the slab, carrying sand and dust aloft to form dark spots (c). Dark material is transported downwind, forming fans (d). Because of the near equality in sublimation rates between spot and spot-free regions, the entire upper surface lowers evenly (d). The high-velocity gas flow at the base of the ice erodes the spiders whose centre lies beneath a spot (d). Eventually the ice is completely removed and the dark granular material is back on the surface (e).

the order of 10 m s^{-1} at 10 m from the vent (Fig. 4c). The spots are granular materials that have been carried to the top of the ice by these jets. The dark material in spots and fans is relatively coarse ($>100 \mu\text{m}$), initially falls near the vent, is $<1 \text{ mm}$ thick, and is subsequently transported across the surface by the wind to form fans (Fig. 4d).

The small differences in albedo between the dark spot material on the surface (~ 0.23) and the dark substrate beneath the ice (~ 0.24) results in nearly equal solar energy input and nearly equal sublimation in spot and spot-free regions. In the spots the dark sand absorbs sunlight, resulting in sublimation at the surface, whereas over the rest of the cap the sublimation occurs at the base. As a result of this near equality in sublimation rates, the dark spots do not burrow down to the substrate, but instead the entire upper surface lowers evenly (Fig. 4d). The high-velocity gas flow at the base of the ice erodes the 'spiders' whose centres lie beneath vents^{7,19}. Once established, this spider morphology is responsible for the year-to-year repeatability of the spot locations. Eventually the ice is completely removed and the dark granular material is back on the surface (Fig. 4e). Finally, the formation of some spots the following year before sunrise is probably due to residual subsurface summer heat, with the rapid increase in number after sunrise due to solar input that begins the process again.

Near-infrared spectra from Mars Express Observatoire pour la Mineralogie l'Eau, les Glaces et l'Activite (OMEGA)²³ support several elements of this jetting model. CO₂ ice is indeed observed in cryptic terrain near "Manhattan" from Ls 197° to Ls 242°. The observed spectra at Ls 197° require significant dust contamination at the surface of the CO₂ ice layer. At the time of this first OMEGA observation the vents and spots had already been active for about

30 days, and in many cases have reached their final configuration (Fig. 3a–c). Between Ls 197 and 223° this surface became contaminated by a thin veneer of dust (7 wt%; 0.5–1 mm)²³, in excellent agreement with the THEMIS observations and the model presented here; virtually all spots and fans have formed by Ls 223° (Fig. 3d). Furthermore, the dust contamination seen by OMEGA probably formed by the same venting process that formed the spots. Late in the season sand and dust accumulation may reduce the sunlight penetration into the ice²³, consistent with the reduction in spot activity following Ls 229° (Fig. 3). However, this surface dust layer is strongly scattering²³ and some of the obscuration of the CO₂ spectra bands may be by water ice²⁴, with the result that a significant fraction of the solar energy probably penetrates into the slab even during the waning phase of ice sublimation after Ls 229°.

The regional brightening and disappearance of some fans at Manhattan could be due to the condensation of a thin veneer of bright water ice and/or dust at several times throughout the spring. Poleward transport of water from the edge of the retreating cap has been suggested to account for the observed annulus of brightening near the retreating cap edge, and may be the source of these surface frosts. Areas of exposed water ice^{14,25} around the south polar cap may provide an additional potential source of water.

The formation of translucent slab ice, basal sublimation and cap levitation, high-velocity CO₂ gas jetting, and eruptions of sand are unlike any phenomena observed on Earth, although a related process of solar-powered geysers has been proposed for Triton^{26,27}. Each year, this process will winnow out the finest-grained materials that can be carried away by the wind in suspension, resulting in a layer of material that has been well sorted by a unique aeolian process that operates vertically.

The erosion and vertical stirring of surface materials under seasonal slab ice may have significantly altered the metre-scale sedimentary structures in the polar-layered deposits in a manner similar to bioturbation on the Earth. This erosion and redeposition of the surface material on vertical scales of a few metres may have produced sedimentary structures that reflect this modification process, rather than the initial depositional environment. If so, this process may present major complications to the interpretation of the sedimentary record observed in upcoming Polar Lander observations, and must be considered in relating this record to the climate history of Mars.

Received 4 April; accepted 30 May 2006.

1. Malin, M. C. & Edgett, K. S. Mars Global Surveyor Mars Orbiter Camera: Interplanetary cruise through primary mission. *J. Geophys. Res.* **106**, 23429–23570 (2001).
2. Cantor, B., Malin, M. & Edgett, K. S. Multiyear Mars Orbiter Camera (MOC) observations of repeated Martian weather phenomena during the northern summer season. *J. Geophys. Res.* **107**, doi:10.1029/2001JE001588 (2002).
3. Edgett, K. S., Supulver, K. D. & Malin, M. C. Spring defrosting of martian polar regions: Mars Global Surveyor MOC and TES monitoring of the Richardson crater dune field, 1999–2000. In *2nd International Conference on Mars Polar Science and Exploration (University of Iceland, Reykjavik, 21–25 August 2000)* 34 (Univ. Iceland, Reykjavik, 2000).
4. Kieffer, H. H., Titus, T., Mullins, K. & Christensen, P. R. Mars south polar cap behaviour observed by TES: Seasonal cap evolution controlled by frost grain size. *J. Geophys. Res.* **105**, 9653–9700 (2000).
5. Malin, M. C. & Edgett, K. S. Early defrosting of the 1999 south polar seasonal frost cap: Evidence of interannual climate change? *Lunar Planet. Sci. [CD-ROM]* XXXI, abstr. 1052 (2000).
6. Kieffer, H. H. Annual punctuated CO₂ slab-ice and jets on Mars. In *2nd International Conference on Mars Polar Science and Exploration (University of Iceland, Reykjavik, 21–25 August 2000)* 93 (Univ. Iceland, Reykjavik, 2000).
7. Piqueux, S., Byrne, S. & Richardson, M. I. Sublimation of Mars's southern seasonal CO₂ cap and the formation of spiders. *J. Geophys. Res.* **108**, 5084, doi:10.1029/2002JE002007 (2003).
8. Supulver, K. D., Edgett, K. S. & Malin, M. C. Seasonal changes in frost cover in the martian south polar regions: Mars Global Surveyor MOC and TES monitoring of the Richardson Crater dune field. *Lunar Planet. Sci.* XXXII, abstr. 1966 (2001).
9. Kossacki, K. J. & Leliwa-Kopystynski, J. Non-uniform seasonal defrosting of subpolar dune field on Mars. *Icarus* **168**, 201–204 (2004).

10. Christensen, P. R. *et al.* The Thermal Emission Imaging System (THEMIS) for the Mars 2001 Odyssey Mission. *Space Sci. Rev.* **110**, 85–130 (2004).
11. Christensen, P. R. *et al.* Morphology and composition of the surface of Mars: Mars Odyssey THEMIS results. *Science* **300**, 2056–2061 (2003).
12. Christensen, P. R., Gorelick, N. S., Mehall, G. L. & Murray, K. C. THEMIS Public Data Releases. Planetary Data System node. (<http://themis-data.asu.edu>) (Arizona State Univ., Tempe, Arizona, 2006).
13. Christensen, P. R., Kieffer, H. H. & Titus, T. Infrared and visible observations of south polar spots and fans. *Eos (Fall Meeting Suppl.)* **86**(52), abstr. P23C–04 (2005).
14. Titus, T. N., Kieffer, H. H. & Christensen, P. R. Exposed water ice discovered near the south pole of Mars. *Science* **299**, 1048–1051 (2003).
15. Smith, D. E., Zuber, M. T. & Neumann, G. A. Seasonal variations of snow depth on Mars. *Science* **294**, 2141–2146 (2001).
16. Feldman, W. C. *et al.* CO₂ frost cap thickness on Mars during northern winter and spring. *J. Geophys. Res.* **108**, 5103, doi:10.1029/2003JE002101 (2003).
17. Boynton, W. V. *et al.* Distribution of hydrogen in the near surface of Mars: Evidence for subsurface ice deposits. *Science* **297**, 81–85 (2002).
18. Titus, T. N., Kieffer, H. H., Mullins, K. F. & Christensen, P. R. TES Pre-mapping data: Slab ice and snow flurries in the Martian north polar night. *J. Geophys. Res.* **106**, 23181–23196 (2001).
19. Kieffer, H. H. Behaviour of solid CO₂ on Mars: A real zoo. *6th Int. Conf. Mars (Pasadena, California, 20–25 July 2003)* abstr. 3158 (California Institute of Technology, Pasadena, 2003).
20. Kieffer, H. H. & Titus, T. N. TES mapping of Mars' north seasonal cap. *Icarus* **154**, 162–189 (2001).
21. Mitrofanov, I. G. *et al.* CO₂ snow depth and subsurface water-ice abundance in the northern hemisphere of Mars. *Science* **27**, 2081–2084 (2003).
22. Haberle, R. M., Mattingly, B. & Titus, T. N. Reconciling different observations of CO₂ ice mass loading of the martian north polar cap. *Geophys. Res. Lett.* **31**, L05702, doi:10.1029/2004GL019445 (2004).
23. Langevin, Y. *et al.* No signature of clear CO₂ ice from the 'cryptic' regions in Mars' south seasonal polar cap. *Nature* doi:10.1038/nature05012 (this issue).
24. Kieffer, H. H. Interpretation of the martian polar cap spectra. *J. Geophys. Res.* **75**, 510–514 (1970).
25. Bibring, J.-P. *et al.* Perennial water ice identified in the south polar cap of Mars. *Nature* **428**, 627–630 (2004).
26. Soderblom, L. A. *et al.* Triton's geyser-like plumes—Discovery and basic characterization. *Science* **250**, 410–415 (1990).
27. Kirk, R. L., Soderblom, L. A. & Brown, R. H. Subsurface energy storage and transport for solar-powered geysers on Triton. *Science* **250**, 424–429 (1990).
28. Christensen, P. R. *et al.* The Mars Global Surveyor Thermal Emission Spectrometer experiment: Investigation description and surface science results. *J. Geophys. Res.* **106**, 23823–23871 (2001).

Acknowledgements We thank the members of the Odyssey and THEMIS mission teams and the ASU and USGS software development group for their efforts in obtaining and processing the THEMIS data used in this work.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to P.R.C. (phil.christensen@asu.edu).

Spontaneous skyrmion ground states in magnetic metals

U. K. Rößler¹, A. N. Bogdanov^{1,2} & C. Pfleiderer^{2,3,4}

Since the 1950s, Heisenberg and others have addressed the problem of how to explain the appearance of countable particles in continuous fields¹. Stable localized field configurations were searched for an ingredient for a general field theory of elementary particles, but the majority of nonlinear field models were unable to predict them. As an exception, Skyrme succeeded in describing nuclear particles as localized states, so-called 'skyrmions'². Skyrmions are a characteristic of nonlinear continuum models ranging from microscopic to cosmological scales^{3–6}. Skyrmionic states have been found under non-equilibrium conditions, or when stabilized by external fields or the proliferation of topological defects. Examples are Turing patterns in classical liquids⁷, spin textures in quantum Hall magnets⁸, or the blue phases in liquid crystals⁹. However, it has generally been assumed that skyrmions cannot form spontaneous ground states, such as ferromagnetic or antiferromagnetic order, in magnetic materials. Here, we show theoretically that this assumption is wrong and that skyrmion textures may form spontaneously in condensed-matter systems with chiral interactions without the assistance of external fields or the proliferation of defects. We show this within a phenomenological continuum model based on a few material-specific parameters that can be determined experimentally. Our model has a condition not considered before: we allow for softened amplitude variations of the magnetization, characteristic of, for instance, metallic magnets. Our model implies that spontaneous skyrmion lattice ground states may exist generally in a large number of materials, notably at surfaces and in thin films, as well as in bulk compounds, where a lack of space inversion symmetry leads to chiral interactions.

Hobart and Derrick established with general arguments that particle-like configurations are not stable in the majority of nonlinear field models^{10,11}. However, a few exceptions have been found. Skyrme showed that particle-like excitations of continuous fields exist in the presence of certain nonlinear interactions². As a drawback, the interactions considered by Skyrme are physically not transparent, because they involve higher-order derivative terms that are technically intractable. Therefore, Skyrme's approach is not viable in the context of simple ordered states in condensed matter that are ruled by short-range interactions. In contrast, a physically transparent exception to the Derrick–Hobart theorem has been recognized in systems with broken inversion symmetry, where chiral interactions lead to skyrmion excitations in condensed matter systems^{12–14}. Chiral interactions exist in many different systems, such as (1) spin-orbit interactions in non-centrosymmetric materials, also referred to as Dzyaloshinsky–Moriya interactions¹⁵, (2) in non-centrosymmetric ferroelectrics, (3) for certain structural phase transitions, (4) in chiral liquid crystals, and (5) in the form of Chern–Simons terms in gauge field theories (see section V of the Supplementary Information or

<http://xxx.lanl.gov/abs/cond-mat/0603104> for references). Yet, it was concluded that spontaneous skyrmion ground states are not stable, and can only be induced by an appropriate applied field¹³ or the proliferation of topological defects⁹. Finally, a different mechanism has been found in quantum Hall ferromagnets. Here, charge, as an additional degree of freedom, allows us to stabilize skyrmions in an external field. As for the other examples, skyrmions in quantum Hall ferromagnets do not form spontaneous ground states.

We studied magnetic skyrmions that are driven by chiral interactions, because the magnetism of condensed-matter systems provides perhaps the richest setting of physics problems in which different kinds of chiral interactions exist. In a phenomenological quasi-classical continuum approximation we considered systems with a hierarchy of well-separated energy and length scales, where ferromagnetic exchange favouring spin alignment is the strongest scale, followed by chiral interactions favouring spin rotations on intermediate scales, and magnetic anisotropies and dipolar interactions that determine favourable directions of spins as the weakest scale. The model hence specifically addresses intermediate length scales, that is, it applies to modulations with typical lengths of a few hundred up to several thousand angstroms observed in many chiral helix magnets^{14–16}. The magnetic free energy density is written in the form:

$$f = Am^2 \sum_{ij} (\partial_i n_j)^2 + \eta A (\nabla m)^2 + f_D(\mathbf{m}) + f_0(m) \quad (1)$$

where the first and second term in A and ηA (where $\eta, A > 0$), respectively, describe the magnetic stiffness, with $\mathbf{n}$ being a unit vector along the magnetization, $\mathbf{m} = m\mathbf{n}$. As discussed in detail in the Supplementary Information, the gradient terms describe additional softening of the amplitude of the magnetization for $\eta < 1$. Values of η for EuS, Ni and MnSi of 0.925, 0.65 and 0.4, respectively, underscore the importance of our predictions in real materials. The third term $f_D(\mathbf{m}) \equiv D\mathcal{L}(\mathbf{m})$ in the free energy describes the chiral interactions, where the Dzyaloshinsky–Moriya constant D determines its handedness and strength. The Dzyaloshinsky–Moriya interactions can be expressed in terms of so-called Lifshitz invariants: that is, linear, antisymmetric gradient terms of the magnetization $\mathcal{L}_{ij}^{(k)} = (m_i \partial_k m_j - m_j \partial_k m_i)$, where i, j, k are combinations of cartesian coordinates x, y, z that are consistent with the symmetry class of the system under consideration. The fourth term $f_0(m)$ includes nongradient terms of the free energy and may be expanded according to Landau theory in even powers of m . We neglect weaker contributions such as magnetic anisotropies or dipolar interactions, and address their importance in the discussion of our results.

Expanding the contribution $f_0(m)$ according to Landau theory for small values of m as $f_0 = a(T - T_C)m^2 + bm^4 + \dots$ leads, in the

¹IFW Dresden, PO Box 270116, D-01171 Dresden, Germany. ²Physikalisches Institut, Universität Karlsruhe, Wolfgang-Gaede-Str. 1, D-76128 Karlsruhe, Germany.

³Physik Department E21, Technische Universität München, James-Frank-Strasse, D-85748 Garching, Germany. ⁴Forschungszentrum Karlsruhe, Institut für Festkörperphysik, D-76021 Karlsruhe, Germany.

absence of chiral interactions ($f_D = 0$), to the conventional Curie temperature T_C of centrosymmetric systems. Thus, when the temperature drops below T_C , the energy density of a ferromagnetically spin-aligned state is lowest. Chiral interactions ($f_D < 0$) favouring rotations of the moments with respect to each other reduce the energy density further. This stabilization of twisted magnetic states occurs only by the competition of the Dzyaloshinsky–Moriya interactions with the exchange, expressed by the form of the gradient energy. Therefore, the exact form of $f_0(m)$ is not decisive for the stability of skyrmions. In particular a form of $f_0(m)$ that includes higher-order terms with $b < 0$ and a stabilizing term m^6 for a first-order transition will change only quantitative relations and details of the phase diagram. Possibly, it may lead to discontinuous nucleation of magnetized states. Figure 1 illustrates the twisted magnetic structures with rotating magnetization and the associated energy densities. For rotations along a single given direction, shown in Fig. 1a, the reduction in energy density, shown in Fig. 1d, is uniform. The transition to the ferromagnetic state is consequently preempted by a transition to a one-dimensionally modulated state at a temperature:

$$T_h = T_C + \frac{D^2}{4Aa} \quad (2)$$

As shown in Fig. 1b and d, rotations of the magnetization in two dimensions reduce the energy density even further near its nucleation point at the centre of the cylindrical structure $\rho = 0$. Considering only the energy reduction at $\rho = 0$, we obtain an upper limit for the temperature T_D for the onset of this so-called double-twisting:

$$T_D = T_C + \frac{D^2}{2Aa} \quad (3)$$

The ratio A/D measures the pitch of the chiral modulations in the ordered helical state, while D/a measures the chirality of magnetic fluctuations in the paramagnetic state. Hence, D/a is related to a chiral

component of the susceptibility. The parameters a and b are the initial susceptibility and mode coupling parameter, respectively. This permits estimates of T_h and T_D from experiments.

As shown in Fig. 1d, the energy reduction due to a double-twist is no longer uniform. In fact, for fixed magnetization amplitude the energy reduction turns into an excess of the energy density at larger distances ρ from the centre, which outweighs the initial reduction. This destabilizes the double-twist structure altogether, which is the reason why magnetic skyrmions prior to the work reported here were believed to be unstable. Skyrmion textures may only form spontaneously when no additional energy is required for double-twist rotations at large ρ . The latter may be achieved by permitting the magnitude of the moment to be soft. As shown in Fig. 1c and 1d, skyrmions under these conditions have the lowest energy density, when the moment decreases with increasing distance from the centre.

The results of a comprehensive numerical analysis of the free energy are summarized in Figs 2 and 3 (see equation (1) and the Methods section). This analysis establishes rigorously that stable cylindrical skyrmions form spontaneously in a finite temperature interval $T_t < T < T_s$ of the phase diagram Fig. 2, where the limits T_h and T_D are also marked. The magnetic structure of the skyrmions is characterized by a magnetization vector $\mathbf{m} = (m, \theta, \psi)$ that rotates outward from the axis in all directions, while the modulus $m(\rho)$ shrinks and decreases towards the boundary of the tube at radius R (Fig. 1c). The amplitude of the local magnetization $m(\rho)$ vanishes continuously when approaching T_s from below, as shown in Fig. 3. Thus, the skyrmion nucleation and the formation of the skyrmion texture coincide. The increase of the magnetization with decreasing temperature drives the transition from the skyrmion phase to the uniform helix phase at a lower temperature T_t . The transition into a one-dimensionally modulated state by transforming the structure in Fig. 1b or in Fig. 1c into the helix Fig. 1a cannot be performed by a smooth deformation of the skyrmions, because these two textures have different topologies. For this reason, a first-order transition is generally expected at a temperature T_t . In the phase diagram, the transition line T_t has been determined from the approximate condition that the average energy of the circular skyrmion cell is equal to that of the helix. The skyrmion phase vanishes as $T_s \rightarrow T_t$ in the limit $\eta \rightarrow 1$.

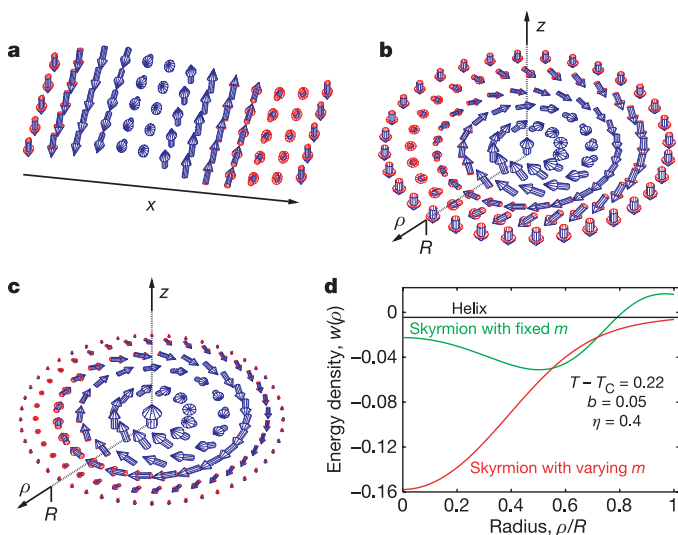


Figure 1 | Three chiral modulated structures for noncentrosymmetric ferromagnets and comparison of their energy density. **a**, One dimensional modulated structure propagating along a direction x with constant modulus of magnetic moments. **b**, Circular cross-section through a skyrmionic excitation with fixed modulus. The modulation has propagation axes in all directions ρ from the centre of a cylindrical tube. **c**, A circular skyrmion cell. The magnetization has maximum modulus m at the skyrmion axis, $\rho = 0$, and equals zero at the cell boundary R . Magnetic structures **a**, **b** and **c** can exist in noncentrosymmetric crystals of class T and of classes D_n (ref. 13), in particular in ferromagnets with B20-structure such as MnSi (ref. 16). **d**, Comparison for the local energy density per unit volume for the helical modulation (**a**) and skyrmions with fixed and varying magnetic moment m (**b** and **c**), along a radial direction ρ .

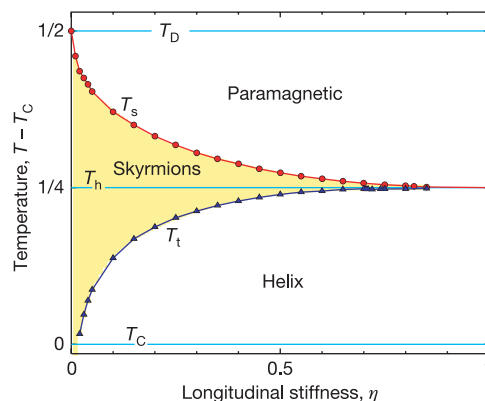


Figure 2 | Phase diagram of a chiral ferromagnet in terms of temperature versus longitudinal stiffness parameter (T, η). The diagram is based on the stability of the cylindrical skyrmions shown in Fig. 1c. Above the ferromagnetic Curie temperature T_C two characteristic temperatures, equation (3), rule the phase diagram: T_D for the instability against double twisting and T_h for the formation of the helical phase. The skyrmion phase is stable between the line $T_s(\eta)$ for the continuous transition into the paramagnetic phase and the line $T_t(\eta)$ for the first-order transition into the helix. The helical phase exists as a metastable phase between T_h and $T_t(\eta)$. The first-order line $T_t(\eta)$ has been calculated for models with $b = 0.05$.

In a chiral magnet, the formation of an extended skyrmion condensate is now subject to: (1) weak residual magnetic anisotropies and dipolar interactions in the crystalline environment, and (2) the magnetization in the regime joining the skyrmion tubes. Under the most favourable conditions, the skyrmion condensate may form a regular lattice in analogy to the vortex lattice of type-II superconductors. In fact, there exists a fundamental analogy between magnetic skyrmion lattices and Abrikosov lattices. The formation of stable spatially modulated states in superconductors and in magnets can be justified by the concept of negative domain wall energies^{14,17,18}.

We illustrate the stability of a skyrmion condensate for this case by the full solution for an extended state in an isotropic system in two dimensions. We note that the solution as shown in Fig. 4 applies also to three-dimensional systems, when the solutions are artificially constrained to remain homogeneous in the third spatial direction (perpendicular to the plane shown in Fig. 4). The structure of this solution shows in particular that the core region of the skyrmions is unaffected, as stated above, while the magnetization between the skyrmions mediating their interactions does not even need to be strictly zero for the skyrmion lattice to become the ground state.

The spontaneous formation of skyrmions applies to all symmetry classes without inversion centre, because the modulus of the magnetization m and the polar angle θ —which control the energetic stability of the skyrmions, and are given by equations (7) and (8) in the Methods section, respectively—are identical for all relevant symmetry classes (see refs 12 and 14 for a discussion). In contrast, the behaviour of the azimuthal angle ψ affects only the manner of rotation of the magnetic moments in the radial direction. General considerations of the stability of localized textures for chiral magnets suggest that droplet-like skyrmions are not stable in three dimensions¹². Therefore, we believe that the two-dimensionally localized skyrmion tubes discussed so far will form extended ground-state condensates in three-dimensional systems. However, in a three-

dimensional environment bending and terminating skyrmion tubes imposes a cost in energy that determines the nature of the condensate. We therefore expect that the skyrmion condensate in isotropic three-dimensional materials may have a liquid-like or amorphous appearance driven by the frustration between preferred orientations and their elastic interactions in the condensed structure.

Our study was inspired by the recent discovery of a non-Fermi liquid phase and partial magnetic order in the cubic itinerant-electron helimagnet MnSi (refs 19, 20). A very detailed discussion of the possible relevance of the predicted spontaneous skyrmion ground state in MnSi is given in the Supplementary Information. We note that recent theories on the partial magnetic order in MnSi by Tewari *et al.*²¹ and Binz *et al.*²² also suggest inhomogeneous twisted states with modulated amplitude, however without identifying the basic mechanism of stable skyrmions in such textures. More importantly, the elementary mechanism for a ferromagnet (illustrated in

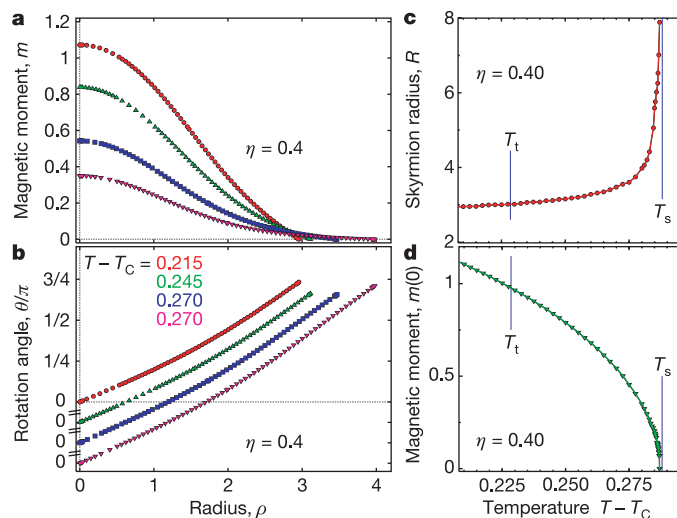


Figure 3 | Details of skyrmion solutions for $\eta = 0.4$ and $b = 0.05$ at various temperatures. (See equations (7) and (8) in Methods section.) **a**, Variation of the modulus of the magnetization as function of radius at different temperatures. **b**, Variation of the angle of rotation, θ , as a function of the radius at different temperatures. At low temperatures the magnetization rotates from angle zero in the centre to angles $\theta(R) < \pi$ at the cell boundaries. The angles $\theta(R)$ depend on temperature. **c**, Evolution of the radius of the skyrmion cell up to the transition into the paramagnetic state. In the transition region both the radius of the skyrmion cell R and the angles $\theta(R)$ grow unlimitedly by approaching the critical temperature T_s . **d**, Corresponding evolution for the amplitude magnetization in the centre $m(0)$. The modulus gradually decreases to zero as the temperature approaches the critical value T_s while the radius of the skyrmions diverges.

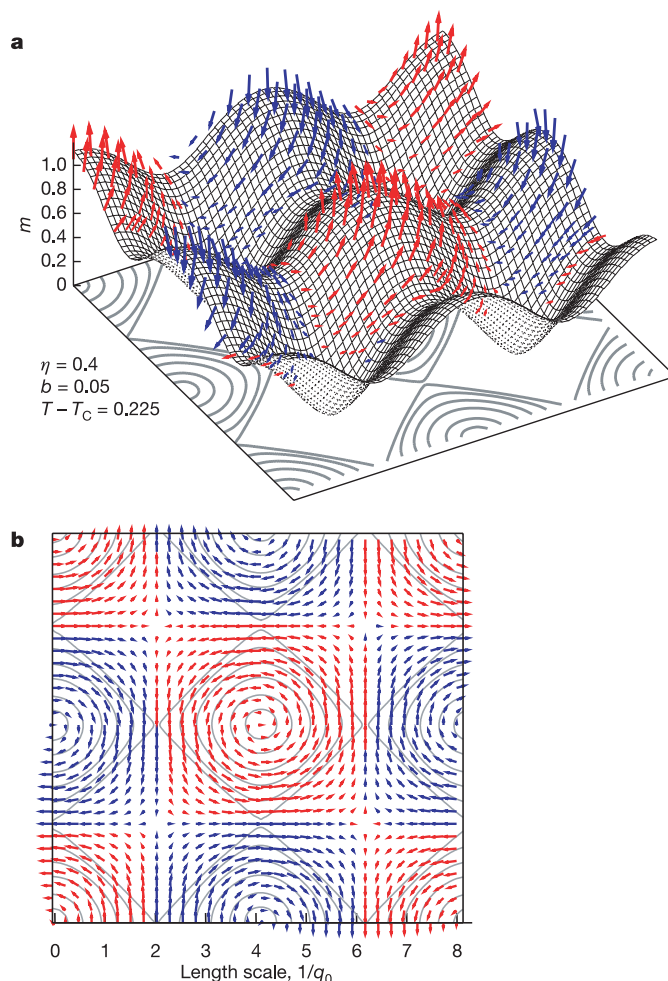


Figure 4 | Structure of a two-dimensional skyrmion lattice, derived as a minimum energy solution for the model's equation (1) with Dzyaloshinskii-Moriya interactions: equation (4). **a**, Overview, showing the modulus m and the corresponding magnetization vectors. **b**, Projection of the magnetization vectors into the base plane. Red/blue signals local magnetization direction with positive/negative components. Contour lines stand for constant modulus, $m = \text{constant}$, in the skyrmion cores. The result applies to thin magnetic films with broken inversion symmetry made from isotropic or cubic (crystallographic class T) metallic materials. A three-dimensional analogue of this dense texture composed of an amorphous arrangement of cylindrical skyrmion strings is consistent with recent experiments in the cubic B20 metals FeGe and MnSi (refs 20 and 29).

Fig. 1) applies to any chiral system with a tendency to form twisted structures. In particular, our model also applies to antiferromagnetic crystals with chiral symmetry by replacing the magnetization with a staggered antiferromagnetic vector order parameter²³. Owing to the broken inversion symmetry at the surfaces of magnetic materials^{24,25}, thin films and interfaces are the largest class of condensed-matter systems in which skyrmion condensates may form spontaneously. Appropriate metallic films of this kind have not been investigated in detail in the regime of interest, for which the predicted appearance of a possible spontaneous skyrmion ground state is shown in Fig. 4.

For a periodic skyrmion lattice neutron scattering will show long-range order akin to the vortex lattice in superconductors, while the scattering intensity of an amorphous arrangement will be akin to partial order in liquid crystals²⁶. In the bulk properties subtle evidence for the predicted intermediate phase is expected, when going from paramagnetism to helical magnetic order (quantitative examples for the size of the temperature interval are given in the Supplementary Information). Real-space imaging of magnetic textures by magnetic force microscopy, spin-sensitive tunnelling, magnetic X-ray microscopy, and Lorentz transmission electron microscopy are finally expected to provide evidence of skyrmion lattices akin to vortex lattices in the Shubnikov phase of superconductors. The helical structure in the chiral magnet (Fe, Co)Si was recently imaged directly²⁷, because of the exceptionally long helix period in this material, but the technique may introduce uncontrolled strains that interfere with the intrinsic properties. It is to be expected that the resolution of improved imaging techniques will soon be sufficient to take pictures of the skyrmions predicted here.

Our calculations shed new light on a seminal result by Wright and Mermin⁹, which is believed to prove the absence of skyrmion ground states akin to the blue phases of liquid crystals in chiral ferromagnets^{9,28}. However, the argument by Wright and Mermin is limited to $\eta = 1$, so our results for $\eta < 1$ establish the possible existence of blue phases in chiral ferromagnets. Whereas the blue phases in chiral liquid crystals are stabilized by topological defects in the form of disclination lines, which do not exist in magnetic systems, skyrmion textures in chiral magnets emerge instead as a condensate of strictly localized skyrmions with lines or sheets of vanishing magnetic order between them (Fig. 4). Fluctuations and the competition between different frustrated couplings entail the possibility of internal transitions between various textures with liquid, glassy or lattice-like organization of skyrmions. Thus, the predicted magnetic skyrmion textures promise rich phase diagrams, similar to those of liquid crystals and Abrikosov lattices of flux-lines in type-II superconductors¹⁷.

METHODS

Cubic noncentrosymmetric magnets (for example, MnSi and other B20 compounds¹⁶) belong to the crystallographic class *T*. For this class the Dzyaloshinsky–Moriya energy is

$$f_D = D(\mathcal{L}_{yx}^{(z)} + \mathcal{L}_{xz}^{(y)} + \mathcal{L}_{zy}^{(x)}) \equiv D\mathbf{m} \times \text{rot}(\mathbf{m}) \quad (4)$$

where D is the Dzyaloshinsky constant¹⁶. The helix solutions for these systems are described by a linear wave with a wave number $q_0 = D/(2A)$ and constant magnetization amplitude $m^2 = 2a(T_h - T)/b$ (refs 9, 16). For all uniaxial systems the invariants f_D are listed in ref. 14. In particular, for uniaxial magnets with C_{nv} symmetry ($n = 3, 4, 6$) we have $f_D = D(\mathcal{L}_{xz}^{(x)} - \mathcal{L}_{yz}^{(y)})$. The n -fold axis is directed along z . Examples are Tb₃Al₂ or Dy₃Al₂ with tetragonal noncentrosymmetric structure C_{4v} (ref. 13). The structure of the cylindrically symmetric skyrmions is given by solutions of the Euler equations for the free energy equation (1). The magnetization vector $\mathbf{m} = (m, \theta, \psi)$ is aligned along the cylinder or z axis in the skyrmion centre, $\rho = 0$. In cubic and uniaxial systems solutions for a skyrmion homogeneous along the z direction are given by $m(\rho)$, $\theta(\rho)$, $\psi = \text{sign}\phi + \phi_0$, where $\text{sign} = +$ or $-$ and ϕ_0 is a constant, both determined by the specific symmetry class^{13,14}. The dependence $m(\rho)$ and $\theta(\rho)$ describes the variation of the amplitude of the magnetic moments and their angle as function of the distance ρ to the cylinder axis in cylindrical coordinates,

respectively. After substitution of the solutions for $\psi(\phi)$ and an integration with respect to z and ϕ the energy of the skyrmion is reduced to the following functional $\mathcal{F}_s = 2\pi L \int_0^\infty f_s(m, \theta) \rho d\rho$ with energy density:

$$f_s = \eta A \left(\frac{dm}{d\rho} \right)^2 + A \mathcal{G}(\theta, \theta_\rho) m^2 + q_0^{-2} f_0(m) \quad (5)$$

$$\mathcal{G}(\theta, \theta_\rho) = \left(\frac{d\theta}{d\rho} \right)^2 + \frac{\sin^2 \theta}{\rho^2} - 2 \left[\left(\frac{d\theta}{d\rho} \right) + \frac{\sin \theta \cos \theta}{\rho} \right] \quad (6)$$

where L is the length of the skyrmion along the z direction, and the spatial coordinate ρ is measured in q_0^{-1} units. The Euler equations for the functional (6):

$$\eta \left[\left(\frac{d^2 m}{d\rho^2} \right) + \frac{1}{\rho} \left(\frac{dm}{d\rho} \right) \right] - A \mathcal{G}(\theta, \theta_\rho) m - q_0^{-2} \frac{\partial f_0}{\partial m} = 0 \quad (7)$$

$$m \left[\left(\frac{d^2 \theta}{d\rho^2} \right) + \frac{1}{\rho} \left(\frac{d\theta}{d\rho} \right) - \frac{\sin \theta \cos \theta}{\rho^2} - \frac{2 \sin^2 \theta}{\rho} \right] + 2 \left[\left(\frac{d\theta}{d\rho} \right) - 1 \right] \left(\frac{dm}{d\rho} \right) = 0 \quad (8)$$

with the boundary conditions $\theta(0) = 0$, $m(0) = m_0$, $m(R) = 0$, describe the structure of the isolated skyrmion and yield the energy as a function of m_0 and R . The equilibrium skyrmion profiles are derived by minimization of $\mathcal{F}_s(m_0)$ with respect to m_0 and R ; the radially integrated energy densities of the vortices per lengths (Fig. 1d) are given by $w(\rho) = (2/\rho^2) \int_0^\rho f_s(m, \theta) \rho d\rho$. Full solutions for extended two-dimensional models have been constructed numerically using a finite-difference scheme on rectangular lattices with dynamically optimized lattice parameters and periodic boundary conditions. The search for ground-states and checks on their stability were done using a continuous variable Monte Carlo simulated annealing method and mesh-refinement.

The unit length used throughout this paper is $1/q_0$, which is the inverse of the twisting length in chiral magnets with Dzyaloshinsky–Moriya couplings. Energy and temperature are given in equivalent units of $D^2/(2Aa)$ as in equations (2) and (3).

Received 5 March; accepted 6 July 2006.

- Heisenberg, W. *Einführung in die einheitliche Feldtheorie der Elementarteilchen* (Hirzel, Stuttgart, 1967).
- Skyrme, T. H. A non-linear field theory. *Proc. R. Soc. Lond. Ser. A* **260**, 127–138 (1961).
- Durrer, R., Kunz, M. & Melchiorri, A. Cosmic structure formation with topological defects. *Phys. Rep.* **364**, 1–81 (2002).
- Bäuerle, C. et al. Laboratory simulation of cosmic string formation in the early Universe using superfluid ³He. *Nature* **382**, 332–334 (1996).
- Brey, L., Fertig, H. A., Cote, R. & MacDonald, A. H. Skyrme crystal in a two-dimensional electron gas. *Phys. Rev. Lett.* **75**, 2562–2565 (1995).
- Al'Khawaja, U. & Stoof, H. T. C. Skyrmions in a ferromagnetic Bose-Einstein condensate. *Nature* **411**, 918–920 (2001).
- Cross, M. C. & Hohenberg, P. C. Pattern formation outside of equilibrium. *Rev. Mod. Phys.* **65**, 852–1112 (1993).
- Sondhi, S. L., Karlhede, A., Kivelson, S. A. & Rezayi, E. H. Skyrmions and the crossover from the integer to the fractional quantum Hall effect at small Zeeman energies. *Phys. Rev. B* **47**, 16419–19426 (1993).
- Wright, D. C. & Mermin, N. D. Crystalline liquids—the blue phases. *Rev. Mod. Phys.* **61**, 385–432 (1989).
- Hobart, R. H. On instability of a class of unitary field models. *Proc. Phys. Soc. Lond.* **82**, 201–202 (1963).
- Derrick, G. H. Comments on nonlinear wave equations as models for elementary particles. *J. Math. Phys.* **5**, 1252–1253 (1964).
- Bogdanov, A. New localized solutions of the nonlinear field-equations. *JETP Lett.* **62**, 247–251 (1995).
- Bogdanov, A. & Hubert, A. Thermodynamically stable magnetic vortex states in magnetic crystals. *J. Magn. Magn. Mater.* **138**, 255–269 (1994).
- Bogdanov, A. N. & Yablonskii, D. A. Thermodynamically stable 'vortices' in magnetically ordered crystals. The mixed state of magnets. *Sov. Phys. JETP* **68**, 101–103 (1989).
- Dzyaloshinskii, I. E. Theory of helicoidal structures in antiferromagnets. *Sov. Phys. JETP* **19**, 960–971 (1964).
- Bak, P. & Jensen, M. H. Theory of helical magnetic-structures and phase-transitions in MnSi and FeGe. *J. Phys. C* **13**, L881–L885 (1980).
- Abrikosov, A. A. On the magnetic properties of superconductors of the second group. *Sov. Phys. JETP* **5**, 1174–1183 (1957).
- Bogdanov, A. N., Rößler, U. K. & Pfleiderer, C. Modulated and localized structures in cubic helimagnets. *Physica B* **359–361**, 1162–1164 (2005).
- Pfleiderer, C., Julian, S. R. & Lonzarich, G. G. Non-Fermi-liquid nature of the normal state of itinerant-electron ferromagnets. *Nature* **414**, 427–430 (2001).
- Pfleiderer, C. et al. Partial order in the non-Fermi-liquid phase of MnSi. *Nature* **427**, 227–231 (2004).
- Tewari, S., Belitz, D. & Kirkpatrick, T. R. Blue quantum fog: chiral condensation in quantum helimagnets. *Phys. Rev. Lett.* **96**, 047207 (2006).

22. Binz, B., Vishwanath, A. & Aji, V. Theory of the helical spin crystal: a candidate for the partially ordered state of MnSi. *Phys. Rev. Lett.* **96**, 207202 (2006).
23. Bogdanov, A. N., Rößler, U. K., Wolf, M. & Müller, K.-H. Magnetic structures and reorientation transitions in noncentrosymmetric uniaxial antiferromagnets. *Phys. Rev. B* **66**, 214410 (2002).
24. Fert, A. Metallic multilayers. in *Materials Science Forum* (eds Chamberod, A. & Hillairat, J.) **59–60**, 439 (1990).
25. Bogdanov, A. N. & Rößler, U. K. Chiral symmetry breaking in magnetic thin films and multilayers. *Phys. Rev. Lett.* **87**, 037203 (2001).
26. Chaikin, P. M. & Lubensky, T. C. *Principles of Condensed Matter Physics* Ch. 2 (Cambridge Univ. Press, Cambridge, UK, 1995).
27. Uchida, M., Onose, Y., Matsui, Y. & Tokura, Y. Real-space observation of helical spin order. *Science* **311**, 359–361 (2006).
28. Hornreich, R. M., Kugler, M. & Shtrikman, S. Localized instabilities and the order-disorder transition in cholesteric liquid crystals. *Phys. Rev. Lett.* **48**, 1404–1407 (1982).
29. Lebech, B. in *Recent Advances in Magnetism of Transition Metal Compounds* (eds Kotani, A. & Suzuki, N.) 167–178 (World Scientific, Singapore, 1993).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The permanent address of A.N.B. is the Donetsk Institute for Physics and Technology, Donetsk, Ukraine. We wish to thank B. Binz, H. Eschrig, I. Fischer, A. Möbius, A. Rosch, H. von Löhneysen, M. Vojta and W. Zwerger for support and discussions. We are particularly grateful to P. Böni for discussions on his studies of longitudinal magnetic fluctuations in EuS, Ni and MnSi. A.N.B. thanks the DFG-Graduiertenkolleg GRK 284 'Kollektive Pänomene im Festkörper' for financial support. C.P. acknowledges support in the form of a Helmholtz-Hochschul-Nachwuchsgruppe at the Universität Karlsruhe in the initial part of this project.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to U.K.R. (u.roessler@ifw-dresden.de).

LETTERS

Resonant slow fault slip in subduction zones forced by climatic load stress

Anthony R. Lowry^{1†}

Global Positioning System (GPS) measurements at subduction plate boundaries often record fault movements similar to earthquakes but much slower, occurring over timescales of ~ 1 week to ~ 1 year. These 'slow slip events' have been observed in Japan^{1,2}, Cascadia^{3–7}, Mexico^{8,9}, Alaska¹⁰ and New Zealand¹¹. The phenomenon is poorly understood, but several observations hint at the processes underlying slow slip. Although slip itself is silent, seismic instruments often record coincident low-amplitude tremor in a narrow (1–5 cycles per second) frequency range¹². Also, modelling of GPS data^{3,7,9} and estimates of tremor location¹³ indicate that slip focuses near the transition from unstable ('stick-slip') to stable friction at the deep limit of the earthquake-producing seismogenic zone. Perhaps most intriguingly, slow slip is periodic at several locations, with recurrence varying from 6 to 18 months depending on which subduction zone (or even segment) is examined^{4–6,9}. Here I show that such periodic slow fault slip may be a resonant response to climate-driven stress perturbations. Fault slip resonance helps to explain why slip events are periodic, why periods differ from place to place, and why slip focuses near the base of the seismogenic zone. Resonant slip should initiate within the rupture zone of future great earthquakes, suggesting that slow slip may illuminate fault properties that control earthquake slip.

Observations of slow slip behaviour and associated tremor have spurred inquiry into several possible mechanisms. Seismic observations of harmonic tremor suggest a role for fluid flux, as similar volcanic tremor results from flux-driven resonance of fluid-filled cracks¹⁴. Coincidence of tremor and slip has prompted the inference that de-watering of subducted oceanic lithosphere reduces effective normal stress and lubricates slow slip¹⁵. However, tremor source locations are distributed throughout the upper plate, suggesting instead that tremor is triggered by stress/strain during slip¹³.

Focusing of slip near the base of the seismogenic zone indicates a role for frictional dynamics. GPS data are ambiguous about deformation sources because the signal-to-noise ratios are low; data from the Cocos–North America plate boundary (Fig. 1) are fitted about equally well by models that locate slip partly within the zone of interplate earthquake rupture and by models that place slip entirely downdip⁹ (Fig. 1b, c). Tremor source locations are less uncertain, and Cascadia sources occupy a narrow horizontal band¹³ that approximately overlies the frictional transition from seismogenic velocity-weakening ('stick-slip') to velocity-strengthening ('stable')¹⁶. Numerical models of rate- and state-dependent frictional slip provide independent support for this interpretation. Models specifying a Ruina–Dieterich constitutive law for frictional slip produce slow slip events near the base of the seismogenic zone^{17,18}. However, frequency and propagation characteristics of these simulated events are quite different from those observed in the Earth.

Slow slip in the Puget basin of Cascadia recurs at intervals of

14.7 ± 1.2 months (refs 4, 7), and predictions of event timing have been reliable enough for experiments to be designed around them. This quasi-periodicity prompted Shen *et al.*¹⁹ to propose that slip is modulated by the climate-driven pole-tide, which has period $T \approx 14.3$ months (the pole-tide is a gravitational effect of oscillations in the Earth's rotational axis excited by ocean loading). They modelled plate boundary stress due to the pole-tide and found small (up to ~ 160 Pa peak-to-peak) stress variations with peak shear at or just after the Puget basin events, and they also qualitatively argue that transitional rate-state rheological conditions favour slip at such long periods.

Slow slip elsewhere recurs at 6- to 18-month intervals, depending on location. Events in Guerrero, Mexico, are quasi-periodic with a 12.0 ± 0.3 month period⁹ (Fig. 1). In addition to the 14.7-month Puget basin behaviour, Cascadia events recur at 10.9 ± 1.2 months in northern California⁵ and at 14 months in the Explorer plate region, where slip is ~ 6 months out of phase with Puget activity⁶. Events in the Shikoku region of Japan occur once every six months with alternating initiation locations and directions of propagation², suggesting annual periodicity with a 6-month phase difference at either end of the segment.

The ubiquity of quasi-periodic behaviour suggests that cyclical processes at 6–18-month periods somehow induce fault slip. In this Letter, I examine 12-month periodic slow slip in Guerrero, and propose fault system resonance as a quantitative mechanism for amplification of slip response to small cyclical stress perturbations. The Earth's response to loading by climatic redistribution of the atmosphere, hydrosphere and cryosphere is an obvious candidate for inducing stress at periods corresponding to slow slip. From equations for conservation of momentum, Poisson's equation, elasticity and boundary conditions used to calculate load Love numbers²⁰, I derive expressions applicable everywhere in the Earth's interior for the six independent stress tensor components generated by surface loading (provided as Supplementary Information). Figure 2a depicts an example history of stress at a point on the subduction thrust beneath GPS station COYU. Fault shear and normal stress are shown for the period 1995–2005, assuming a continental PREM²¹ Earth model and a Climate Prediction Center model of global surface hydrologic mass²². The latter was chosen because continental hydrology dominates global deformation at the annual period²³ that is the focus here. Also shown in Fig. 2b are peak-to-peak shear stress changes everywhere on the plate interface during the same period.

These calculations demonstrate several important features of environmental stress on the plate interface. First, shear and normal stress in Fig. 2a are in phase, and although a few of the Guerrero slip events nearly coincide with peak shear, peak slip-rate follows peak stress by 2.9 ± 1.5 months. The phase of hydrologic load stress depends on location, but phase on the Cocos–North America plate interface varies by less than a month. Second, shear stress

¹Department of Physics, University of Colorado, Boulder, Colorado 80309-0390, USA. [†]Present address: Department of Geology, Utah State University, Logan, Utah 84322, USA.

perturbations on the subduction thrust are largest (and normal stress perturbations smallest) near the coastline, coincident with the base of the seismogenic zone (compare with Fig. 1b, c). The amplitude of the stress change is partly conditioned by orographic focusing of precipitation in the coastal Sierra Madre del Sur and partly by the increase of horizontal stress with depth. However, shear and normal stress are most sensitive to fault dip, and the unusual shallow-steep-shallow dip geometry of the plate interface in Guerrero (Fig. 2b inset) amplifies shear and dampens normal stress on the steeply dipping section.

Hydrologic stress perturbations are a factor of two or more larger than those predicted for the pole-tide¹⁹, but 300 Pa is still a negligible fraction ($\sim 0.01\%$) of the stress drop during large interplate earthquakes on the Cocos–North America plate boundary²⁴. Hence, it is necessary to assess whether climate-driven stress is at all significant relative to tectonic stress accumulation on similar timescales. Figure 2c depicts the tectonic shear stress accumulation rate in Guerrero (described further in Methods). Peak-to-peak hydrologic stress (Fig. 2b) ranges from $\sim 1\%$ of annual tectonic accumulation (Fig. 2c) near the trench to 5–20% at the base of the seismogenic zone, and can exceed tectonic rates at greater depths where the latter drops near zero.

Frictional slip initiates when shear stress τ exceeds $\mu_f \sigma_e$, where μ_f is the coefficient of static friction and $\sigma_e = \sigma_n - P_f$ is effective normal stress, σ_n is true normal stress and P_f is pore fluid pressure. If μ_f is invariant and stress has a small ratio of annual to secular variation like that estimated for southern Mexico, it is difficult to imagine how this physical formalism could organize slip that is quasi-periodic and has the phase properties observed in Fig. 2a. Hence, for fault slip to be forced by environmental stress, one requires (1) that slip must respond more robustly to stress at periods of about one year than to secular stress or stress at other periods, and (2) that the phase of peak slip can shift significantly from that of peak stress.

Slider-block models of slip, given a rate- and state-dependent friction law^{25,26}, suggest a mechanism by which both of these conditions may be met. Rate- and state-dependent friction varies,

for example, as the Dieterich law²⁷, $\mu_f = \mu_0 + a \ln(V/V^*) + b \ln(V^* \Theta / D_c)$, in which μ_0 , a , b and D_c are experimentally determined material constants, V is slip velocity, V^* is a normalizing velocity and Θ is a time-dependent state variable. Perfettini *et al.*²⁵ find that slip response to a given stress can amplify dramatically within a narrow band of resonant periods, provided that frictional conditions are velocity-weakening (as defined by $b - a > 0$) and slip is stabilized by the fault's elastic stiffness $k = \dot{\tau} / \dot{u} > k_c$, where u is slip deficit and

$$k_c = \frac{\sigma_n(b-a)}{D_c} \quad (1)$$

is a critical stiffness below which slip destabilizes. In that case, excitation by stress at period T with shear stress amplitude and normal stress amplitude $\bar{\tau}$ and $\bar{\sigma}_n$, respectively, perturbs the slip velocity by an amount approximated by²⁶:

$$\tilde{V} = q V_{\text{rpm}} \frac{q[\bar{\sigma}_n(\mu_{ss} - \alpha) - \bar{\tau}] - i(\bar{\sigma}_n \mu_{ss} - \bar{\tau})}{k D_c - a \sigma_n q^2 + i q D_c (k - k_c)} \quad (2)$$

in which the dimensionless pulsation $q = 2\pi D_c / T V_{\text{rpm}}$, V_{rpm} is relative plate motion velocity, μ_{ss} is steady-state friction and α describes frictional response to a change in normal stress. Slip is unbounded (that is, resonant) when the denominator goes to zero; at critical stiffness ($k = k_c$) this occurs at a critical period:

$$T_c = \frac{2\pi D_c}{V_{\text{rpm}}} \sqrt{\frac{a}{b-a}} \quad (3)$$

Fig. 3a depicts an example calculation of perturbational slip velocity $\tilde{V}$ as a function of stiffness k and period T using equation (2). The parameters used (lower-left inset) are consistent with expectations near the base of the seismogenic zone in southern Mexico. Slip response amplifies significantly for a narrow range of periods near T_c when stiffness $k = k_c$, and at shorter periods when $k > k_c$. In addition to stiffness, the resonant period depends on local values of a , b , D_c and V_{rpm} . At periods other than the resonant period, slip amplitude is sensitive to assumed model parameters and particularly

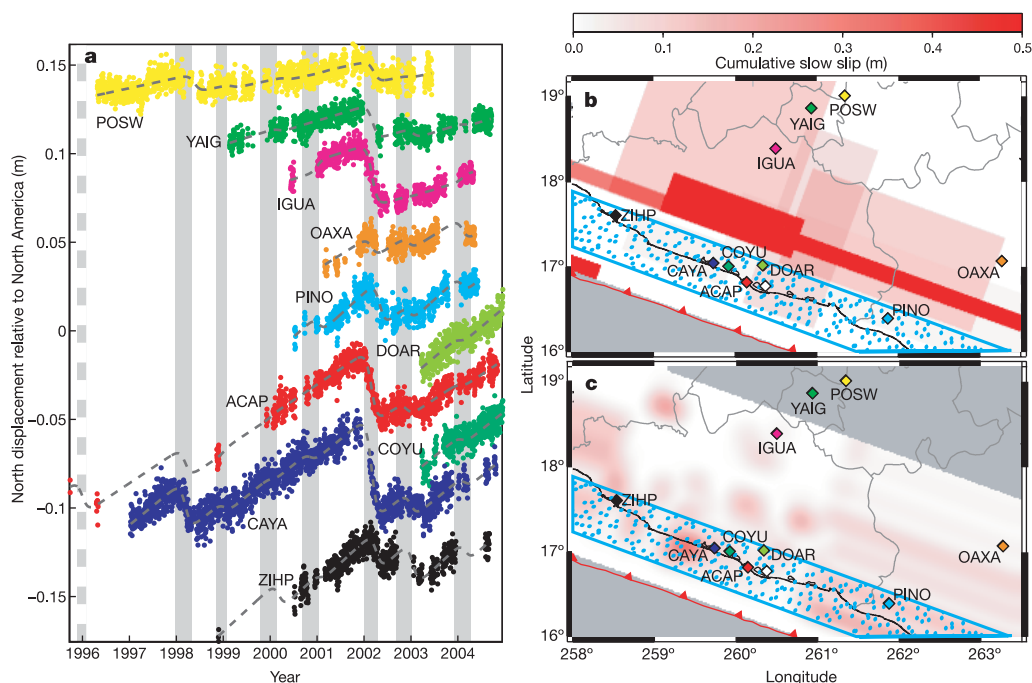


Figure 1 | GPS time series and slow slip models from Guerrero, Mexico. **a**, GPS north positions. Filled circles are daily coordinates, dashed lines are best-fit model⁹. Grey bands indicate timing of slow slip from the GPS data. **b**, **c**, Maps showing locations of GPS sites and surface-projected models of

cumulative slow thrust slip⁹; in **b** each event was modelled as uniform slip in a single rectangular patch; in **c** slip was estimated in discretized elements. The stippled blue region approximates the zone of rupture in past great earthquakes.

to $(b - a)$. The very small $(b - a) = 10^{-8}$ assumed in Fig. 3a, likely to occur only near the top and bottom of the seismogenic zone, amplifies slip response to environmental stress by about two orders of magnitude more than would a value of $(b - a) = 10^{-6}$ (Fig. 3a inset). Using $(b - a) = 0.005$, near the laboratory-derived parameter at middle depths of the seismogenic zone²⁸, reduces response by nearly six orders of magnitude. Hence, slip resonance explains why slow slip should be favoured near the transition from velocity-weakening to velocity-strengthening at the base of the seismogenic zone. Elsewhere in the seismogenic zone, amplification is negligible and/or too narrow-band to excite slip over an area large enough to observe with GPS. In southern Mexico, the largest events slip early on the shallow part of the fault and later on the deeper part⁹, suggesting that slip propagates from the seismogenic zone to greater depth, analogous to postseismic slip excited by earthquakes. Dependence of slip amplitude on $(b - a)$ similarly explains why creep events simulated in numerical models focus at the frictional transition¹³.

Phase differences between peak stress and peak slip (for example, Fig. 2a) also contain information about resonance properties. Figure 3b shows details of slip velocity amplitude near its maximum for several examples of k/k_c , using the same parameterization as in

Fig. 3a. The corresponding dependence of phase is shown in Fig. 3c. In this example, phase γ is zero (that is, slip and stress are in phase) for excitation periods greater than the maximum velocity perturbation, and $\gamma = \pi$ (perfectly out of phase) for periods less than the maximum perturbation. When the resonant period and excitation period coincide, $\gamma = \pi/2$. Phase is somewhat sensitive to parameters assumed in equation (2), but the pattern of $\gamma \approx 0$ on one side of the saddle and $\gamma \approx \pi$ on the other, with transitional phase between, is a persistent feature given small $(b - a)$. Extrapolating to phase relations observed in slow slip events, this implies that the resonant period for the plate boundary in Guerrero is almost exactly one year. The Puget basin⁴ ($\gamma \approx -\pi/4$) and Explorer plate⁶ ($\gamma \approx \pi$) regions of Cascadia apparently resonate at periods slightly offset from (and on opposite sides of) the pole-tide. Similarly, the northeastern ($\gamma \approx 0$)

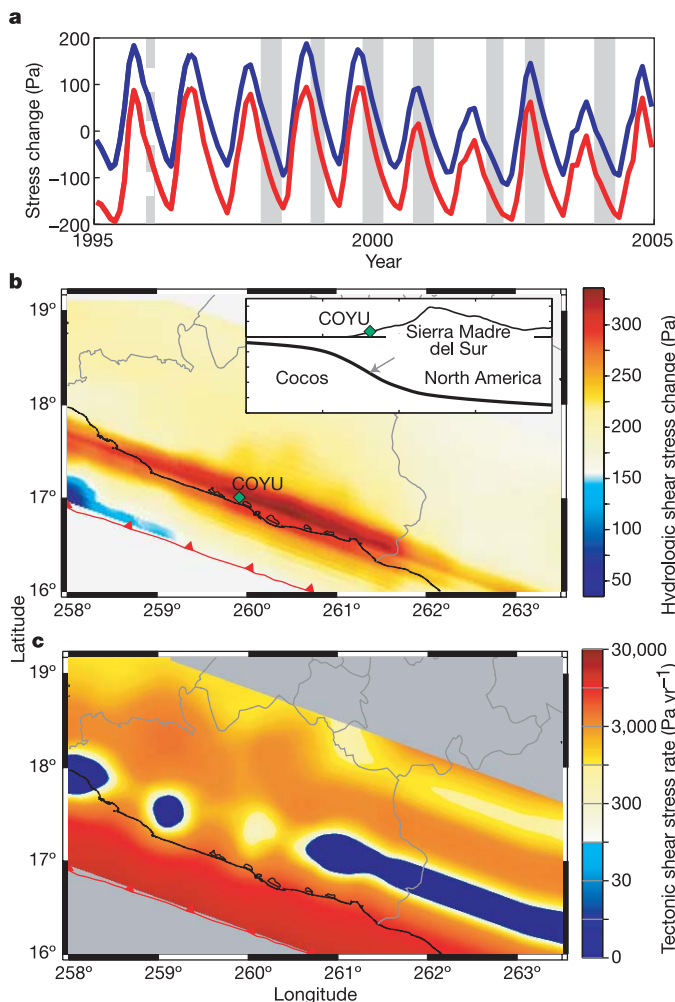


Figure 2 | Stress on the fault surface. **a**, Time series of normal stress (blue; positive indicates fault compression) and shear stress (red; positive favours thrust slip) at a point beneath GPS site COYU. Grey bars denote periods of deep slow slip; peak slip occurs at the centre of the bar. **b**, Map view of peak-to-peak shear stress perturbation, projected from the plate interface to the surface. Inset shows plate geometry and strike-averaged topography versus distance from the trench; arrow indicates location of time series sampled in **a**. **c**, Rate of accumulation of tectonic shear stress.

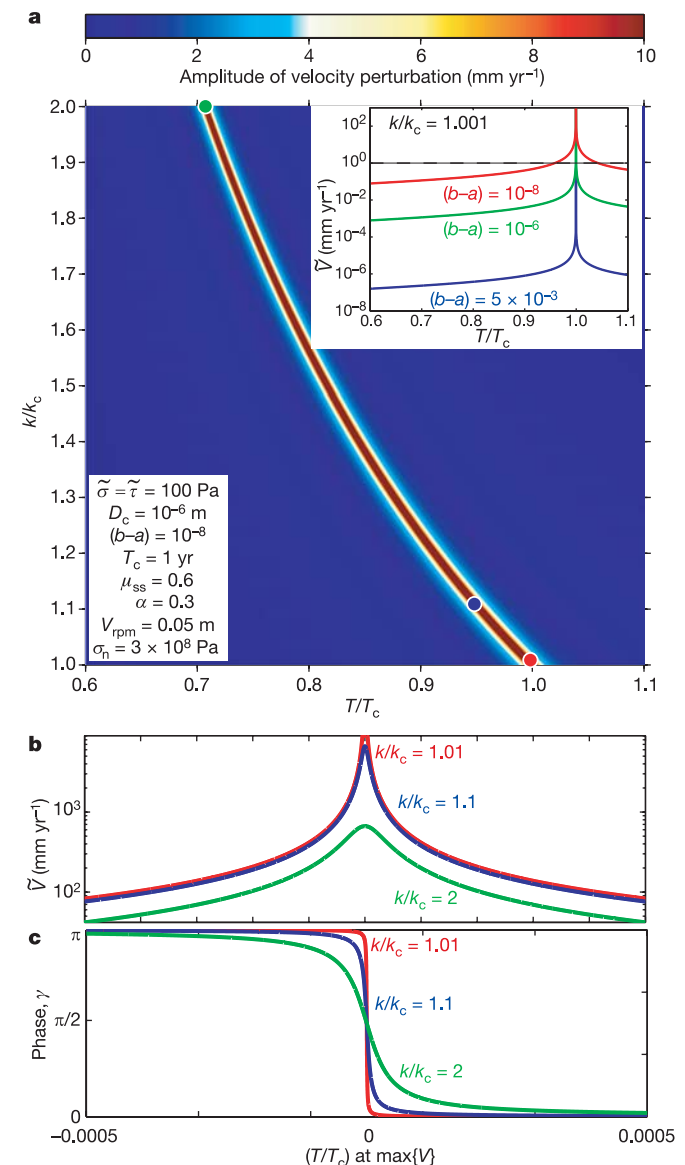


Figure 3 | Relationships of rate- and state-dependent frictional parameters to resonant slow slip. **a**, Slip velocity perturbation as a function of period T and fault stiffness k (normalized to critical values), for frictional parameters shown in lower-left inset (see text for parameter definitions). Upper-right inset, profiles of dependence of velocity amplitude $\tilde{V}$ on T (normalized to critical period, T_c) for alternative values of $(b - a)$. Coloured circles are profiled in detail in **b** and **c**. **b**, Amplitude of slip velocity perturbation near its maximum for $k/k_c = 1.01$ (red), 1.1 (blue) and 2 (green). **c**, Phase of slip (relative peak stress) corresponding to the amplitude profiles shown in **b**.

and southwestern ($\gamma \approx \pi$) Shikoku regions² may resonate at periods slightly offset from (and on opposite sides of) the annual snow loading cycle in Japan.

Critical stiffness and critical period relations may be useful to infer fault frictional properties. Stiffness $k = \dot{\tau}/u$ can be calculated from steady-state slip estimates (for example, using calculations for Fig. 2c). Assuming that Guerrero is near criticality, such that the one-year period $T = T_c$ and $k = k_c$, it is instructive to solve equations (1) and (3) for two unknown frictional parameters a and $(b - a)$ on the fault. Near the base of the Guerrero seismogenic zone, where events most probably initiate, such a calculation yields $a \approx 0.015$ and $(b - a) \approx 10^{-9}$. These values approximate laboratory estimates of frictional parameters at the base of the seismogenic zone²⁸.

When effects of loading in oceans, atmosphere and tides are included, environmental stress is relatively broad-band with a handful of spectral peaks, so there is a rich spectrum of potential forcing signals for fault slip. Resonant fault slip can be observed geodetically if a large enough patch of the fault can resonate at a single period, or can propagate slip downdip into the velocity-strengthening regime. Fault slip resonance explains why slip events should be periodic, why periods differ from place to place, why the phase of peak slip differs from that of peak environmental stress, and why slip focuses near the base of the seismogenic zone. These events must initiate in velocity-weakening conditions—that is, within the zone of nucleation of a future great earthquake—further suggesting that slow slip events can illuminate the frictional properties that control earthquake slip.

METHODS

To determine the relative contribution of tectonic accumulation versus hydrologic load stress, I estimate the annual rate of change of fault stress due to plate boundary deformation. I first calculate stress induced by fault slip using Green's function relations²⁹ applied to steady-state virtual slip on the fault. Virtual slip was modelled on a 20 km mesh discretization of the plate boundary using GPS velocities estimated after subtraction of the displacements during slow slip events⁹. The slip-induced stress distribution was smoothed to remove spurious singularities introduced by discontinuous slip at the dislocation boundaries. I then summed the result with stress rates from a two-dimensional elastic finite element model using the far-field (relative plate motion) kinematics as boundary conditions, to correct for the Green's function assumption of zero displacement at infinite distance. The latter turns out to be a negligibly small correction, amounting to a few tens of Pa yr^{-1} . Figure 2c depicts the resulting estimate of shear stress accumulation rate. Tectonic stress accumulation estimated by this approach is qualitatively similar to stress rates estimated directly from finite element modelling of GPS velocities for Cascadia³⁰. In both cases, shear stress rates in the seismogenic zone are consistent with measured earthquake stress drops when multiplied by the timescale of the seismic cycle, and rates rapidly drop to small values below the depth of rupture in large interplate earthquakes.

Received 24 October 2005; accepted 3 July 2006.

- Hirose, H., Hirahara, K., Kimata, F., Fujii, N. & Miyazaki, S. A slow thrust slip event following the two 1996 Hyuganada earthquakes beneath the Bungo Channel, southwest Japan. *Geophys. Res. Lett.* **26**, 3237–3240 (1999).
- Obara, K., Hirose, H., Yamamizu, F. & Kasahara, K. Episodic slow slip events accompanied by non-volcanic tremors in southwest Japan subduction zone. *Geophys. Res. Lett.* **31**, doi:10.1029/2004GL020848 (2004).
- Dragert, H., Wang, K. & James, T. S. A silent slip event on the deeper Cascadia subduction interface. *Science* **292**, 1525–1528 (2001).
- Miller, M. M., Melbourne, T., Johnson, D. J. & Sumner, W. Q. Periodic slow earthquakes from the Cascadia subduction zone. *Science* **295**, 2423 (2002).
- Szeliga, W., Melbourne, T. I., Miller, M. M. & Santillan, V. M. Southern Cascadia episodic slow earthquakes. *Geophys. Res. Lett.* **31**, doi:10.1029/2004GL020824 (2004).
- Malone, S., Rogers, G., Dragert, H., McCausland, W. & Johnson, D. Review of episodic tremor and slip in Cascadia. *Eos* **85**, abstr. S53A (2004).

- Dragert, H., Wang, K. & Rogers, G. Geodetic and seismic signatures of episodic tremor and slip in the northern Cascadia subduction zone. *Earth Planets Space* **56**, 1143–1150 (2004).
- Lowry, A. R., Larson, K. M., Kostoglodov, V. & Bilham, R. G. Transient slip on the subduction interface in Guerrero, southern Mexico. *Geophys. Res. Lett.* **28**, 3753–3756 (2001).
- Lowry, A. R., Larson, K. M., Kostoglodov, V. & Sanchez, O. The fault slip budget in Guerrero, southern Mexico. *Geophys. J. Int.* (submitted).
- Cohen, S. C. & Freymueller, J. T. Crustal deformation in the southcentral Alaska subduction zone. *Adv. Geophys.* **47**, 1–63 (2004).
- Douglas, A., Beavan, J., Wallace, L. & Townend, J. Slow slip on the northern Hikurangi subduction interface, New Zealand. *Geophys. Res. Lett.* **32**, doi:10.1029/2005GL023607 (2005).
- Rogers, G. & Dragert, H. Episodic tremor and slip on the Cascadia subduction zone: The chatter of silent slip. *Science* **300**, 1942–1943 (2003).
- Kao, H. et al. A wide depth distribution of seismic tremors along the northern Cascadia margin. *Nature* **436**, 841–844 (2005).
- Chouet, B. Resonance of a fluid-driven crack: Radiation properties and implications for the source of long-period events and harmonic tremor. *J. Geophys. Res.* **93**, 4375–4400 (1988).
- Melbourne, T. I. & Webb, F. H. Slow but not quite silent. *Science* **300**, 1886–1887 (2003).
- Flück, P., Hyndman, R. D. & Wang, K. Three-dimensional dislocation model for great earthquakes of the Cascadia subduction zone. *J. Geophys. Res.* **102**, 20539–20550 (1997).
- Shibazaki, B. & Iio, Y. On the physical mechanism of silent slip events along the deeper part of the seismogenic zone. *Geophys. Res. Lett.* **30**, doi:10.1029/2003GL017047 (2003).
- Liu, Y. & Rice, J. R. Aseismic slip transients emerge spontaneously in three-dimensional rate and state modeling of subduction earthquake sequences. *J. Geophys. Res.* **110**, doi:10.1029/2004JB003424 (2005).
- Shen, Z. K., Wang, Q., Bürgmann, R. & Wan, Y. Pole-tide modulation of slow slip events at circum-Pacific subduction zones. *Bull. Seismol. Soc. Am.* **95**, 2009–2015 (2005).
- Wahr, J. M. in *Geodesy and Global Geodynamics* (eds Moritz, H. & Sunkel, H.) 327–380 (Technical Univ. of Graz, Austria, 1983).
- Dziewonski, A. M. & Anderson, D. L. Preliminary reference Earth model. *Phys. Earth Planet. Inter.* **25**, 297–356 (1981).
- Fan, Y. & van den Dool, H. Climate Prediction Center global monthly soil moisture data set at 0.5 degrees resolution for 1948 to present. *J. Geophys. Res.* **109**, doi:10.1029/2004JD004345 (2004).
- Wahr, J. M., Molenaar, M. & Bryan, F. Time variability of the Earth's gravity field: Hydrological and oceanic effects and their possible detection using GRACE. *J. Geophys. Res.* **103**, 30205–30230 (1998).
- Anderson, J. G. et al. Strong ground motion from the Michoacan, Mexico, earthquake. *Science* **233**, 1043–1049 (1986).
- Perfettini, H., Schmittbuhl, J., Rice, J. R. & Cocco, M. Frictional response induced by time-dependent fluctuations of the normal loading. *J. Geophys. Res.* **106**, 13455–13472 (2001).
- Perfettini, H. & Schmittbuhl, J. Periodic loading on a creeping fault: Implications for tides. *Geophys. Res. Lett.* **28**, 435–438 (2001).
- Dieterich, J. H. Modeling of rock friction 1. Experimental results and constitutive equations. *J. Geophys. Res.* **84**, 2161–2168 (1979).
- Blanpied, M. L., Lockner, D. A. & Byerlee, J. D. Frictional slip of granite at hydrothermal conditions. *J. Geophys. Res.* **100**, 13045–13064 (1995).
- Okada, Y. Internal deformation due to shear and tensile faults in a half-space. *Bull. Seismol. Soc. Am.* **82**, 1018–1040 (1992).
- Williams, C. A. & McCaffrey, R. Stress rates in the central Cascadia subduction zone inferred from an elastic plate model. *Geophys. Res. Lett.* **28**, 2125–2128 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements I thank K. Larson for analysis of GPS data used in this paper; V. Kostoglodov, O. Sanchez and J. A. Santiago for contributions to instrumentation and data collection in Mexico; and J. Wahr and R. Bilham for discussions about the topic of this paper. This research was supported by the National Science Foundation.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The author declares no competing financial interests. Correspondence and requests for materials should be addressed to the author (arlowry@himalaya.colorado.edu).

LETTERS

Archaea predominate among ammonia-oxidizing prokaryotes in soils

S. Leininger¹, T. Urich¹, M. Schlöter², L. Schwark³, J. Qi⁴, G. W. Nicol⁵, J. I. Prosser⁵, S. C. Schuster⁴ & C. Schleper¹

Ammonia oxidation is the first step in nitrification, a key process in the global nitrogen cycle that results in the formation of nitrate through microbial activity^{1,2}. The increase in nitrate availability in soils is important for plant nutrition, but it also has considerable impact on groundwater pollution owing to leaching. Here we show that archaeal ammonia oxidizers are more abundant in soils than their well-known bacterial counterparts. We investigated the abundance of the gene encoding a subunit of the key enzyme ammonia monooxygenase (*amoA*) in 12 pristine and agricultural soils of three climatic zones. *amoA* gene copies of Crenarchaeota (Archaea) were up to 3,000-fold more abundant than bacterial *amoA* genes. High amounts of crenarchaeota-specific lipids, including crenarchaeol, correlated with the abundance of archaeal *amoA* gene copies. Furthermore, reverse transcription quantitative PCR studies and complementary DNA analysis using novel cloning-independent pyrosequencing technology demonstrated the activity of the archaea *in situ* and supported the numerical

dominance of archaeal over bacterial ammonia oxidizers. Our results indicate that crenarchaeota may be the most abundant ammonia-oxidizing organisms in soil ecosystems on Earth.

Autotrophic ammonia-oxidizing bacteria (AOB) of the β - and γ -subgroups of proteobacteria have so far been considered the most important contributors to aerobic ammonia oxidation^{3,4}. These organisms usually comprise only a small fraction of the microbiota^{5–7}. Recently, we found genes encoding subunits of a potential ammonia monooxygenase (AMO), the key enzyme of AOB, on a metagenomic soil clone alongside a ribosomal RNA operon of Archaea, affiliated with the phylum Crenarchaeota⁸. After addition of ammonia to bulk soil samples their transcription was induced⁸. Highly similar *amoA* and *amoB* sequences and a linked *amoC* gene were also identified in a metagenomic study of the Sargasso Sea indicating the presence of all three enzyme subunits in crenarchaeota^{9,10}. The predicted existence of ammonia-oxidizing archaea (AOA) was ultimately confirmed by cultivation of a chemolithoauto-

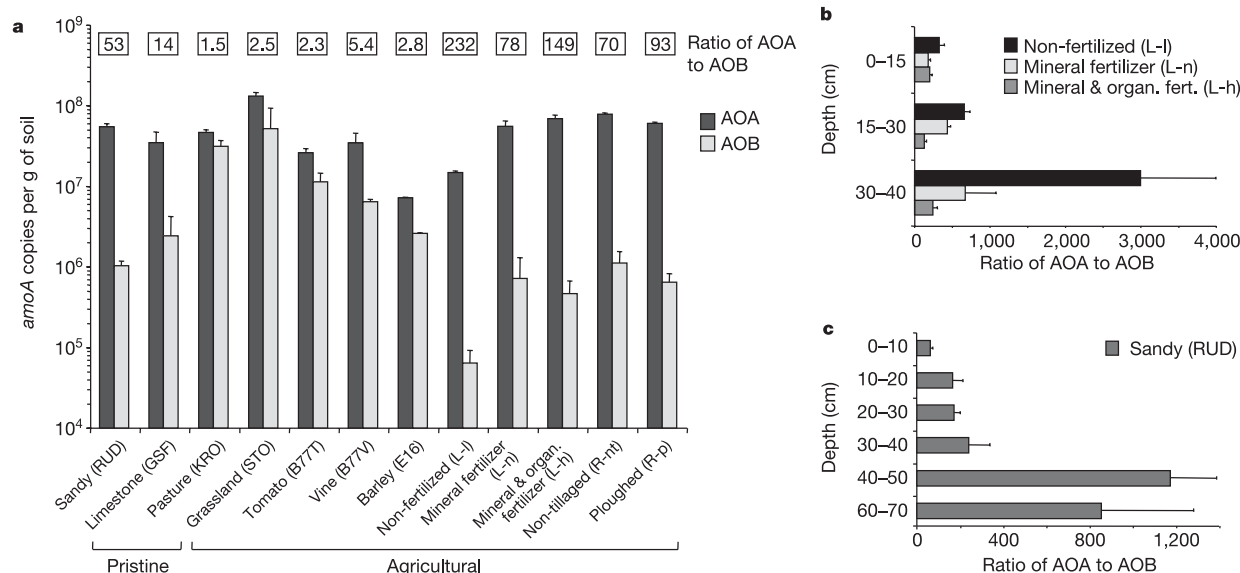


Figure 1 | Archaeal *amoA* genes outnumber bacterial *amoA* genes in topsoils and in deeper layers. **a**, Abundance of AOA and AOB in 12 different topsoils (0–10 cm) expressed as *amoA* copy numbers per g of dried soil, analysed from three independent replicates per site. Ratios of AOA to AOB *amoA* copies are shown in boxes above the chart. **b**, Ratios of AOA to AOB *amoA* gene copies in soil depth profiles of agricultural soils L-I, L-n and

L-h (see Table 1). **c**, Ratios of AOA to AOB *amoA* gene copies in a sandy ecosystem (RUD). The ratios increased with depth in all but one soil (L-h) owing to a dramatic decrease in AOB while AOA stayed constantly high (for absolute numbers see Supplementary Table S2). Error bars indicate standard deviation.

¹Department of Biology, University of Bergen, Jahnebakken 5, N-5020 Bergen, Norway. ²Institute of Soil Ecology, GSF-National Research Center for Environment and Health, Ingolstädter Landstrasse 1, D-85764 Neuherberg, Germany. ³Institute of Geology and Mineralogy, University of Cologne, Zulpicher Strasse 49a, 50674 Cologne, Germany. ⁴Penn State University, Center for Comparative Genomics and Bioinformatics, 310 Wartik Building, University Park, Pennsylvania 16802, USA. ⁵School of Biological Sciences, University of Aberdeen, Cruickshank Building, St Machar Drive, Aberdeen AB24 3UU, UK.

Table 1 | Soil sample characteristics and locations

Name	Soil type	pH	WEON	WEOC	Location
RUD ¹⁸	Sandy grassland	7.1	0.9	7.6	Darmstadt, Germany
GSF	Limestone grassland	6.9	1.08	10.2	Munich, Germany
KRO ²⁴	Pasture land	6.1	3.76	25.85	Bergen, Norway
STO ²⁴	Arable grassland	5.5	1.16	5.5	Bergen, Norway
B77T	Sandy soil, growing tomato	6.0	0.76	5.4	Santorini, Greece
B77V	Sandy soil, growing grapes	6.2	0.87	5.6	Bergen, Norway ²⁴
E16	Sandy soil, growing barley	6.0	0.56	3	Bergen, Norway ²⁴
L-1 ¹⁷	Sandy loam, non-fertilized	6.4	1.22	35.2	Bad Lauchstädt, Germany
L-n ¹⁷	Sandy loam, mineral fertilizer	6.3	2.34	57.1	Bad Lauchstädt, Germany
L-h ¹⁷	Sandy loam, mineral and organic fertilizer	6.7	4.76	85.1	Bad Lauchstädt, Germany
R-nt	Silty clay loam, non-tillaged	7.3	2.3	16.3	Rommersheim, Germany
R-p	Silty clay loam, ploughed	7.3	2.4	7.7	Rommersheim, Germany

Typical soil types from the region of sampling were used, providing a spectrum from sandy soils (Greece) to silty clay loam soils (Rommersheim, Germany). RUD was sampled from 0 to 70 cm depth in 10-cm increments. Other depth profiles were taken from three plots of agricultural soil under different management regimes at Bad Lauchstädt (0–15 cm, 15–30 cm and 30–40 cm). Triplicate soil samples were taken at each site. All soil samples were snap-frozen and stored at -80°C . WEON, water extractable organic nitrogen (mg per kg of dry weight soil); WEOC, water extractable organic carbon (mg per kg of dry weight soil). For more details of soil characteristics, see Supplementary Table S1.

trophic, marine strain which used ammonia as sole energy source and produced nitrite in nearly stoichiometric conversion¹¹. It contained *amoA*, *amoB* and *amoC* genes highly similar to those from the metagenomic soil clone and the marine sequences.

The goal of this study was to quantify ammonia-oxidizing crenarchaeota in terrestrial ecosystems to evaluate their potential ecological impact. For this purpose, we have focused on two biomarkers, the *amoA* gene, which has frequently been used to quantify bacterial ammonia oxidizers^{5–7}, and tetraether lipids diagnostic for crenarchaeota¹². Because the archaeal and bacterial *amoA* genes are sufficiently divergent from each other^{8,13,14}, two specific primer sets were used in real-time (quantitative) polymerase chain reaction (qPCR) to determine the absolute and relative abundance of archaeal and bacterial ammonia oxidizers. Twelve different soils following a sequence from northern to southern Europe were investigated, including natural and managed ecosystems exhibiting different soil types with a wide range of pH, water extractable carbon and nitrogen (for details on soils see Table 1 and Supplementary Fig. S1 and Supplementary Table S1).

In all samples, the archaeal *amoA* copy numbers were high, ranging from 7×10^6 to 1×10^8 per g of dry soil (Fig. 1a, Supplementary Table S1). In contrast, bacterial *amoA* genes varied considerably over three orders of magnitude, similar to earlier findings of AOB abundance in agricultural soils^{5–7,15,16}. Archaeal *amoA* genes dominated over bacterial *amoA* genes in all soils, with ratios ranging from 1.5 to over 230 in topsoils. To analyse the distribution of AOA and AOB at different depths, we investigated an agricultural soil (from Bad Lauchstädt), which has been treated with different amounts and qualities of fertilizers for more than 100 years (ref. 17), and a natural, pristine calcareous grassland soil site (Am Rotböll¹⁸, Fig. 1b and c). Bacterial *amoA* genes in the former declined significantly with depth in the unfertilized and inorganically fertilized sites while archaeal *amoA* genes stayed high, resulting in a maximal AOA to AOB ratio of 3,000. In contrast, both archaeal and bacterial *amoA* copy numbers varied little with depth at the site treated additionally with manure (L-h) and these higher levels of AOA and AOB *amoA* were associated with the highest bioavailability of nitrogen and carbon (Table 1). Interestingly, archaeal *amoA* copy numbers also varied little with depth in the sandy ecosystem (RUD), while the abundance of bacterial *amoA* gene copies again decreased significantly, resulting in an AOA to AOB *amoA* ratio of $>1,000$ at 40 cm (Fig. 1c, Supplementary Table S2). The relative abundance of AOA with respect to the total microbiota varied mostly between 1% and 5% (Supplementary Table S2, assuming arbitrarily an average microbial genome size of 4 megabases, Mb), whereas the highest fraction of AOB in soil microbiota was only 0.23% (STO). To analyse possible differences in community structure of archaea in topsoil and deeper soil horizons, we cloned and sequenced AOA *amoA* from the sandy ecosystem at depths of 0–10 cm and 60–70 cm. A comparison of 36

sequences from each depth shows highly similar *amoA* genes (with 24% maximal sequence divergence at the DNA level), but also clustering, indicating differences in population structure of AOA with depth that might reflect the occurrence of different 'ecotypes' (see tree in Supplementary Fig. S3).

To exclude the possibility that quantification was biased by differences in the sensitivity of the primer sets or the efficiency of the qPCR methods, five different primer combinations for archaeal and bacterial *amoA* and 16S rRNA genes, respectively, were used in a most-probable-number (MPN) PCR experiment (Supplementary Information, Supplementary Table S3 and Supplementary Fig. S4). The results obtained on serial dilutions of two soil DNAs confirmed, with independent primer sets the dominance of AOA over AOB.

Isoprenoid glycerol dialkyl glycerol tetraethers (GDGTs) are unique membrane lipids of archaea and are often used as biomarkers to study their presence and distribution^{19,20}. Among these lipids, crenarchaeol has been identified exclusively in crenarchaeota, particularly from marine environments but also in freshwater sediments and peat bogs^{12,21}. We investigated the amount of archaea-derived isoprenoid GDGT in ten of the soils. Values ranged from 0.04 to $3.24 \mu\text{g}$ per g of soil, of which a significant fraction was crenarchaeol, showing that it occurs in considerable amounts in the soil ecosystems

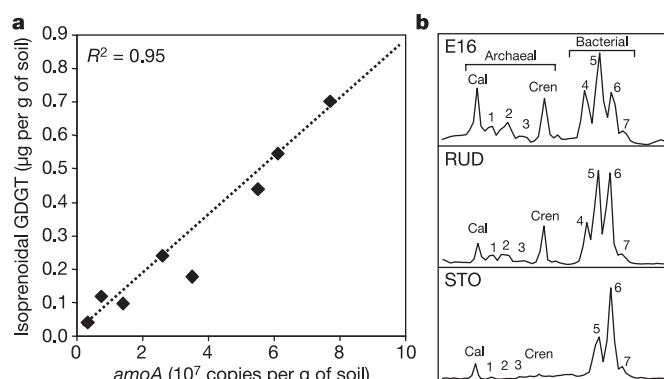


Figure 2 | Isoprenoid tetraether lipids in soils and correlation to *amoA* gene copies. **a**, bivariate plot of GDGT abundance (μg per g of dry soil) versus AOA *amoA* copies (10^7 per g of dry soil). The high coefficient of correlation ($R^2 = 0.95$; $n = 8$) is consistent with ammonia oxidizers constituting a major proportion of crenarchaeota (values from managed pastures KRO and STO excluded). **b**, Chromatographic traces of archaeal and bacterial membrane lipids in three different surface soils (E16, RUD and STO; see Table 1), determined by high-pressure LC-APCI-MS. Isoprenoid GDGTs are attributed to archaea, whereas non-isoprenoid GDGT have been proposed to derive from bacteria²⁶. Cal, caldarchaeol; Cren, crenarchaeol; numbers refer to GDGT structures as displayed in Supplementary Fig. S6.

studied (0.02–0.33 μg per g of soil; Supplementary Tables S1 and S2). A good correlation was found between the amount of GDGT (or crenarchaeol alone) and the abundance of archaeal *amoA* gene copies in eight out of ten soil samples tested, which is consistent with ammonia-oxidizing archaea constituting a significant proportion of crenarchaeota (Fig. 2, and Supplementary Table S1 and Supplementary Fig. S5 for crenarchaeol:*amoA*).

To determine whether *amoA* genes from AOA and AOB populations are actively transcribed in soil and whether transcription correlates with gene abundance, we isolated total RNA from three soils and quantified relative cDNA copies of *amoA* genes after reverse transcription. Remarkably, *amoA* cDNA of both populations was detected in all soils and their ratio correlated well with quantitative DNA measurements; that is, archaeal transcripts dominated in RUD (sandy ecosystem) and KRO (pasture), while almost equal amounts of cDNA copies were detected in STO (grassland) (Fig. 3).

To analyse the activity of AOA and AOB at the community level and to verify their relative abundances by an independent technique, we performed large-scale analysis of a cDNA library using pyrosequencing technology. The cDNA library was obtained from RUD soil in which archaeal 16S rRNA and *amoA* genes were found in approximately equal numbers and archaeal 16S rRNA gene sequences belonged exclusively to crenarchaeota subgroup 1.1b in which the *amo* genes have been detected (ref. 18 and T.U. and C.S., unpublished data). After second-strand synthesis on randomly reverse-transcribed total RNA the resulting double-stranded cDNA was used directly for sequencing without a cloning procedure or PCR amplification²². Out of 210,523 assignable cDNA reads (of a total of 314,041 reads) 1.37% were affiliated with archaea (Supplementary Table S4), a similar proportion to that estimated by *amoA* qPCR of the same sample (1.0% of the microbial cells contained an archaeal *amoA* gene, Supplementary Table S2, RUD 0–10 cm). This is consistent with the assumption that most if not all of the archaea in this soil are capable of ammonia oxidation. While *amoA* cDNA of archaea and bacteria was readily quantified from the same RNA preparations by qPCR (Fig. 3), we found only two reads in the 30 Mb 'pyro-library' that could be assigned to archaeal *amoA* and none was associated with bacterial *amoA* (Supplementary Table S5). However, this is not unexpected given that the vast amount of cDNA appears to be derived from stable ribosomal RNA transcripts (at least 30% of the reads represent 16S rRNA genes, Supplementary Table S4).

To compare further the relative proportions of non-thermophilic

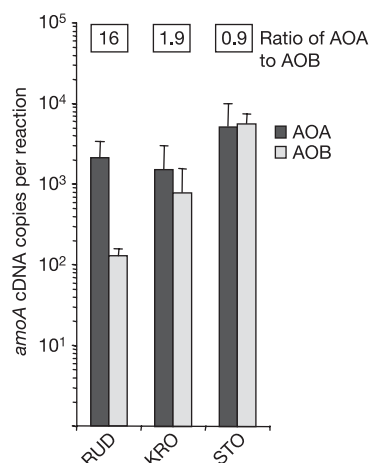


Figure 3 | *amoA* cDNA copies and their AOA:AOB ratios in three different soils. Average of three independent qPCR experiments performed on reverse-transcribed total RNA from fresh soil samples taken in January 2006. Copy numbers of AOA *amoA* transcripts are comparable to those published earlier from the same soil but from a different season⁸. Error bars indicate standard deviation.

crenarchaeota and AOB, we used full-length 16S rRNA genes of both groups to query this database. We identified 570 crenarchaeota- and 37 AOB-assigned sequence reads (Fig. 4a); the ratio of 15:1 is similar to the AOA:AOB ratio determined by *amoA* cDNA copies in the same sample (a ratio of 16:1, see RUD Fig. 3). Similarly, more sequences of crenarchaeota than of AOB were identified when querying the database with a universal probe or with various group-specific probes in different regions of the 16S rRNA gene (Fig. 4b, c and Supplementary Table S5), which is again consistent with dominance of active AOA over AOB in the sample.

In conclusion, our data provide evidence for high abundance of AOA in soils, and extrapolation suggests that they represent the most abundant ammonia-oxidizing organisms in soil ecosystems on Earth. Their high numbers in various ecosystems and at greater soil depths indicate that these organisms are adapted to a broad range of growth conditions and might therefore have a more versatile metabolism than AOB, perhaps being able to grow mixotrophically. Genes predicted to encode components of a modified-3-hydroxypropionate cycle known in carbon-fixing hyperthermophilic crenarchaeota as well as an oxidative tricarboxylic acid cycle have been found in the marine relative, the sponge symbiont *Cenarchaeum symbiosum*²³. This is consistent with both an autotrophic and an organotrophic lifestyle. Furthermore, the genome from *Cenarchaeum* revealed genes encoding homologues of urease and urea transporters beside the *amoA*, *amoB* and *amoC* genes, indicating the potential of AOA to oxidize various reduced nitrogen compounds²³.

Although numerically abundant and transcriptionally active, it remains to be shown whether archaea in soil also dominate with respect to their nitrification activities. It will be important to identify

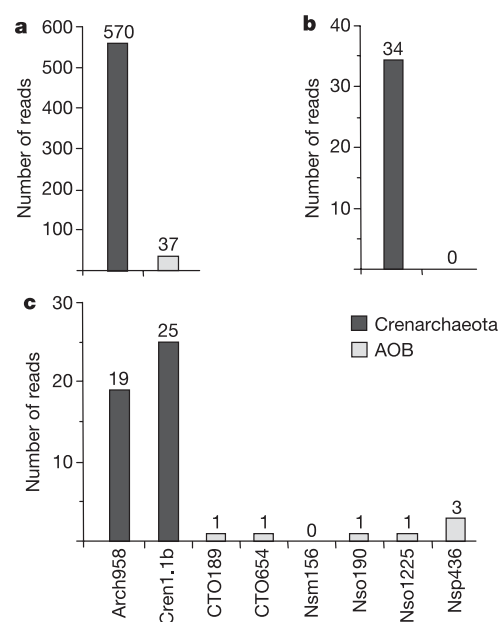


Figure 4 | Identification of rRNA transcripts from crenarchaeota and AOB in 30 Mb of sequence determined from a RUD soil cDNA library (314,000 reads with an average length of 96.4 bp).

a, Search performed by aligning all reads against a set of four full-length crenarchaeal and six full-length AOB 16S rRNA genes, covering the major phylogenetic clusters of these groups, similarity cut-off 97%. **b**, 1,134 sequences that aligned to the universal phylogenetic 16S rRNA probe Uni1392 (ref. 27) were screened for affiliation with AOB or (cren)archaea by keyword searches conducted on individual Blast output files using classification names of Bergey's manual. **c**, (Cren)archaea- and AOB-specific 16S rRNA probes were used in an 'in silico' hybridization experiment, allowing up to one mismatch. All reads identified in **a**, **b** and **c** were verified manually by individual Blast searches against the non-redundant database (see Supplementary Information for more details).

the parameters influencing AOA and AOB populations in soils and to quantify and compare their specific activities under varying environmental conditions. If archaea contribute significantly to nitrification, as their abundance now suggests, estimates of the ecological impact of ammonia oxidation (including greenhouse gas emissions) based on bacterial ammonia-oxidizing activity will need to be re-assessed.

METHODS

Extraction and preparation of nucleic acids. Nucleic acid (RNA and DNA) extractions from all soils were performed using a modification of the method described in ref. 25 after optimization of bead beating times, which led to maximal yield of DNA or of quantified copies of *amoA* in qPCR (see Supplementary Information for details). cDNA was prepared as described previously⁸. DNA concentration was measured by incubating different dilutions of extracted DNA with the fluorescent dye SYBRGreenI (Molecular Probes, see Supplementary Information).

Real-time PCR. For each soil sample, 5 ng of DNA pooled from three replicate samples or 2 µl of cDNA, respectively, was used to quantify the copy numbers of archaeal *amoA* genes and transcripts as described before⁸. The same procedure was applied for *amoA* from bacteria using bacterial *amoA*-specific primers and SYBRGreenI. Both real-time techniques yielded highly reproducible standard curves with fosmid 54d9 (ref. 8) (for archaeal *amoA*) and the cloned *amoA* genes from *Nitrosomonas europaea* ATCC 19718, *Nitrosospira multififormis* ATCC25196, *Nitrosospira* NpAV and *Nitrosovibrio tenuis* NV-12, respectively. All qPCR reactions had high efficiencies and detection limits were as low as ten copies of AOB *amoA* and 30 copies of AOA *amoA* (see Supplementary Information for details). All sample and standard reactions were performed in triplicate and an average value was calculated. Possible inhibitory effects on PCR performance from co-extracted polyphenolic compounds in soil extracts were estimated by three different techniques as described in Supplementary Information. Inhibitory effects were negligible in all samples.

Extraction and quantification of GDGTs. GDGTs were extracted from soil with a polar solvent and extracts were cleaned by Al₂O₃-solid phase extraction (SPE) and filtration through 0.45 µm polytetrafluoroethylene (PTFE) filters. GDGT fractions were analysed by liquid chromatography-atmospheric pressure chemical ionization-mass spectrometry (LC-APCI-MS) on a cyanopropyl (CN)-column and protonated molecular ions were recorded in selected ion monitoring (SIM) as described previously^{21,26} (for details see Supplementary Information).

High-throughput sequencing of cDNA. Total RNA isolated from RUD soil was subjected to reverse transcription with random hexanucleotides and second-strand synthesis (for details see Supplementary Information). The resulting cDNA was sheared into fragments of 100 to 500 base pairs (bp) in length and subsequently used for library construction as described previously²². One run on a 70 × 75 mm PicoTiterPlate format was performed on a GS20 Genome Sequencer (Roche Applied Sciences/454 Life Sciences), yielding 314,041 successful sequencing reads. Sequence data were processed as FastA files for all further data analysis, including megablast or primer searches as described in the Supplementary Information.

Received 11 April; accepted 19 June 2006.

- Kowalchuk, G. A. & Stephen, J. R. Ammonia-oxidizing bacteria: a model for molecular microbial ecology. *Annu. Rev. Microbiol.* **55**, 485–529 (2001).
- Prosser, J. I. & Embley, T. M. Cultivation-based and molecular approaches to characterisation of terrestrial and aquatic nitrifiers. *Antonie Van Leeuwenhoek* **81**, 165–179 (2002).
- Bock, E. & Wagner, M. Oxidation of Inorganic Nitrogen Compounds as an Energy Source. In *The Prokaryotes: An Evolving Electronic Resource for the Microbiological Community* (eds Dworkin, M. et al.) (<http://link.springer-ny.com/link/service/books/10125>) (Springer, New York, 2001).
- Purkhold, U. et al. Phylogeny of all recognized species of ammonia oxidizers based on comparative 16S rRNA and *amoA* sequence analysis: implications for molecular diversity surveys. *Appl. Environ. Microbiol.* **66**, 5368–5382 (2000).
- Hermansson, A. & Lindgren, P. E. Quantification of ammonia-oxidizing bacteria in arable soil by real-time PCR. *Appl. Environ. Microbiol.* **67**, 972–976 (2001).
- Mendum, T. A., Sockett, R. E. & Hirsch, P. R. Use of molecular and isotopic techniques to monitor the response of autotrophic ammonia-oxidizing populations of the beta subdivision of the class proteobacteria in arable soils to nitrogen fertilizer. *Appl. Environ. Microbiol.* **65**, 4155–4162 (1999).
- Phillips, C. J., Paul, E. A. & Prosser, J. I. Quantitative analysis of ammonia oxidising bacteria using competitive PCR. *FEMS Microbiol. Ecol.* **32**, 167–175 (2000).
- Treusch, A. H. et al. Novel genes for nitrite reductase and Amo-related proteins indicate a role of uncultivated mesophilic crenarchaeota in nitrogen cycling. *Environ. Microbiol.* **7**, 1985–1995 (2005).

- Schleper, C., Jurgens, G. & Jonuscheit, M. Genomic studies of uncultivated archaea. *Nature Rev. Microbiol.* **3**, 479–488 (2005).
- Venter, J. C. et al. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74 (2004).
- Konneke, M. et al. Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature* **437**, 543–546 (2005).
- Damste, J. S., Schouten, S., Hopmans, E. C., van Duin, A. C. & Geenevasen, J. A. Crenarchaeol: the characteristic core glycerol dibiphytanyl glycerol tetraether membrane lipid of cosmopolitan pelagic crenarchaeota. *J. Lipid Res.* **43**, 1641–1651 (2002).
- Rotthauwe, J. H., Witzel, K. P. & Liesack, W. The ammonia monooxygenase structural gene *amoA* as a functional marker: molecular fine-scale analysis of natural ammonia-oxidizing populations. *Appl. Environ. Microbiol.* **63**, 4704–4712 (1997).
- Stephen, J. R. et al. Analysis of beta-subgroup proteobacterial ammonia oxidizer populations in soil by denaturing gradient gel electrophoresis analysis and hierarchical phylogenetic probing. *Appl. Environ. Microbiol.* **64**, 2958–2965 (1998).
- Kurola, J., Salkinoja-Salonen, M., Aarnio, T., Hultman, J. & Romantschuk, M. Activity, diversity and population size of ammonia-oxidising bacteria in oil-contaminated landfarming soil. *FEMS Microbiol. Lett.* **250**, 33–38 (2005).
- Okano, Y. et al. Application of real-time PCR to study effects of ammonium on population size of ammonia-oxidizing bacteria in soil. *Appl. Environ. Microbiol.* **70**, 1008–1016 (2004).
- Weigel, A., Russow, R. & Körschens, M. Quantification of airborne N-input in long-term field experiments and its validation through measurements using ¹⁵N isotope dilution. *J. Plant Nutr. Soil Sci.* **163**, 261–265 (2000).
- Ochsenreiter, T., Selezi, D., Quaiser, A., Bonch-Osmolovskaya, L. & Schleper, C. Diversity and abundance of Crenarchaeota in terrestrial habitats studied by 16S RNA surveys and real time PCR. *Environ. Microbiol.* **5**, 787–797 (2003).
- De Rosa, M. & Gambacorta, A. The lipids of Archaeobacteria. *Prog. Lipid Res.* **27**, 153–175 (1988).
- DeLong, E. F. et al. Dibiphytanyl ether lipids in nonthermophilic crenarchaeotes. *Appl. Environ. Microbiol.* **64**, 1133–1138 (1998).
- Hopmans, E. C. et al. A novel proxy for terrestrial organic matter in sediments based on branched and isoprenoid tetraether lipids. *Earth Planet. Sci. Lett.* **224**, 107–116 (2004).
- Margulies, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
- Hallam, S. J. et al. Pathways of carbon assimilation and ammonia oxidation suggested by environmental genomic analyses of marine Crenarchaeota. *PLoS Biol.* **4**, e95 (2006).
- Øvreås, L. & Torsvik, V. V. Microbial diversity and community structure in two different agricultural soil communities. *Microb. Ecol.* **36**, 303–315 (1998).
- Griffiths, R. I., Whiteley, A. S., O'Donnell, A. G. & Bailey, M. J. Rapid method for coextraction of DNA and RNA from natural environments for analysis of ribosomal DNA- and rRNA-based microbial community composition. *Appl. Environ. Microbiol.* **66**, 5488–5491 (2000).
- Weijers, J. W. et al. Membrane lipids of mesophilic anaerobic bacteria thriving in peats have typical archaeal traits. *Environ. Microbiol.* **8**, 648–657 (2006).
- Lane, D. J. et al. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc. Natl Acad. Sci. USA* **82**, 6955–6959 (1985).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank V. Torsvik for use of the qPCR machine, S. L. Jørgensen and V. Torsvik for discussions and L. Knudsen and L. Tomsho for technical assistance. K. Zink is acknowledged for running LC-APCI-MS analyses. We thank for help or support in soil sampling: V. Torsvik (for KRO and STO), E. Schulz (for L-I, L-n and L-h from a long-term field trial in Bad Lauchstädt, Germany), C. Emmerling (for R-nt and R-p), B. Winkler (for GSF), E. Vavoulidou (for B77V, B77T and E16 from the NAGREF Soil Science Institute of Athens). Most of this project was financed through an initial funding of the University of Bergen given to C.S. Part of the pyrosequencing project was paid through the Department of Health using Tobacco Settlement Funds to S.C.S.

Author Contributions The project was conceived and the manuscript was written by C.S., assisted by co-authors. Soil samples were collected and characterized for general parameters by M.S., T.U. and S.L. DNA and RNA extractions were performed by M.S. and S.L. and real-time PCR by S.L.; MPN-PCR and clone libraries were performed by T.U.; GDGT analyses was carried out by L.S.; ds cDNA synthesis and high-throughput sequencing including data analyses was performed by T.U., J.Q. and S.C.S.; and *amoA* phylogeny was performed by G.W.N. and J.I.P.

Author Information Sequences obtained in this study were deposited at GenBank (NCBI) with accession numbers DQ534808–DQ534888. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to C.S. (christa.schleper@bio.uib.no).

LETTERS

Hierarchy and adaptivity in segmenting visual scenes

Eitan Sharon¹, Meirav Galun¹, Dahlia Sharon², Ronen Basri¹ & Achi Brandt¹

Finding salient, coherent regions in images is the basis for many visual tasks, and is especially important for object recognition. Human observers perform this task with ease, relying on a system in which hierarchical processing seems to have a critical role¹. Despite many attempts, computerized algorithms^{2–5} have so far not demonstrated robust segmentation capabilities under general viewing conditions. Here we describe a new, highly efficient approach that determines all salient regions of an image and builds them into a hierarchical structure. Our algorithm, segmentation by weighted aggregation, is derived from algebraic multi-grid solvers for physical systems⁶, and consists of fine-to-coarse pixel aggregation. Aggregates of various sizes, which may or may not overlap, are revealed as salient, without predetermining their number or scale. Results using this algorithm are markedly more accurate and significantly faster (linear in data size) than previous approaches.

We present an algorithm (flow chart in Fig. 1a) that adaptively assembles pixels into small aggregates according to resemblance in luminance. The small aggregates are then assembled in a similar manner, according to resemblance in their properties, into still larger, more complex aggregates. As the aggregates are formed at each level, their statistical properties are accumulated, and their saliency is evaluated according to differences compared with their neighbours⁷. Figure 1b shows a salient segment (III) composed of a hierarchy of aggregates at lower levels (I, II). This increase in both the size and complexity of successive, higher-level aggregates is reminiscent of the primate visual system, in which neurons in successively higher visual areas respond to successively more complex stimulus features, within increasingly larger fields^{8,9}.

Computationally, it is useful to think of segmentation within the framework of cuts in graph theory¹⁰. Each pixel in an image (Fig. 2a) corresponds to a node in a graph (Fig. 2b), coupled to each of its four neighbours according to their similarity in luminance level. The goal is to 'cut' this graph into pieces. A salient segment in the image is one for which the similarity across its border is small, whereas the similarity within the segment is large (for a mathematical description, see Methods). We can thus seek a segment that minimizes the ratio of these two expressions. Despite its conceptual usefulness, minimizing this 'normalized cut' measure is computationally prohibitive, with cost that increases exponentially with image size¹⁰. Approximations to the optimal cut can be obtained using spectral methods, with the most efficient approximation to date having a computational cost proportional to n^3 (ref. 11) (where n is the number of pixels), but this supra-linear cost seems to be yet too demanding for the brain, which deals with very large images.

Moreover, finding salient segments in real images is by nature a problem of multiple scales. Pixel-scale measurements (for example, intensity and colour) alone are insufficient to characterize segments, and larger-scale measurements (for example, average intensity and texture) at multiple scales must be incorporated as well. However, there is an inherent 'chicken-and-egg' difficulty associated with applying coarser measurements to images: when coarse-scale measurements are taken near the boundaries of segments they mix their statistics, and smooth the transition between them (see Supplementary Fig. 1b). Consequently, it is difficult to locate the boundaries between the segments. Ideally, we should apply coarse measurements only within segments, but segment boundaries

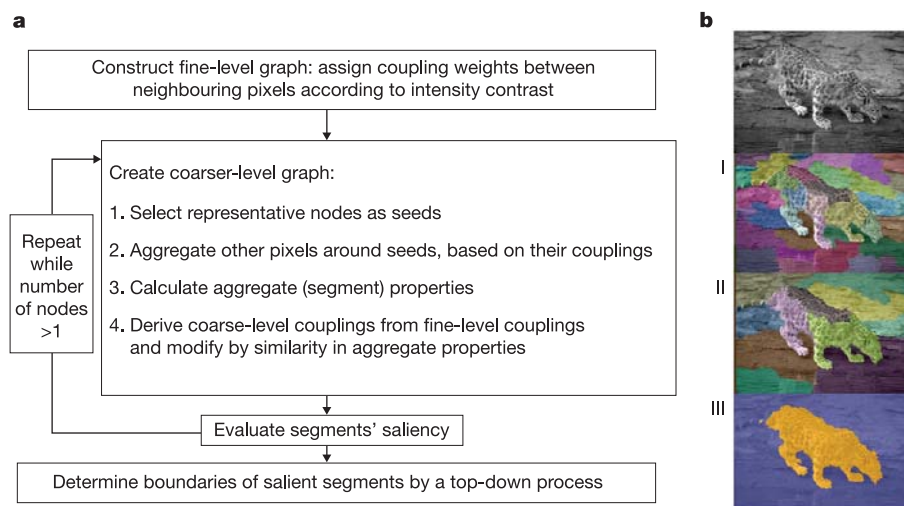


Figure 1 | SWA. **a**, Flow chart. **b**, A hierarchy composing a salient segment and its background. The leopard segment (III) is shown with two out of the ten levels of aggregates composing it (I, II). Original image is shown at the top.

¹Department of Computer Science and Applied Mathematics, The Weizmann Institute of Science, Rehovot 76100, Israel. ²Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital, Charlestown, Massachusetts 02129, USA.

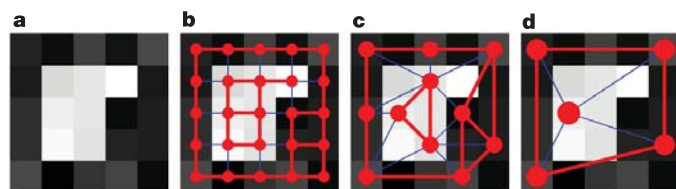


Figure 2 | The multiscale normalized cut graph approach. **a**, A simple image. **b**, Pixels of the image are nodes, represented by filled circles; strong coupling is represented by thick red lines, and weak coupling by thin blue lines. **c**, Adaptive coarsening. Each pixel in **b** is strongly coupled to one of the chosen seeds shown here (thus, pixels strongly coupled to a given seed form an aggregate). Couplings between the seeds are shown. **d**, An additional coarsening level. In this case, this is the level at which the salient segment is detected.

are unknown before the measurements are taken—hence, the chicken-and-egg difficulty. As detailed below, our hierarchical algorithm applies an adaptive coarsening process that solves this problem. (See Supplementary Fig. 1 for an example of the preservation of boundaries by aggregates in our algorithm, compared to regular coarsening.) In addition, the complexity of our algorithm is linear with the number of pixels in the image, on a regular serial computer, and only logarithmic with the image size on a parallel computer. Therefore both the results and computational speed for obtaining them are qualitatively improved under our segmentation by weighted aggregation (SWA) algorithm.

Inspired by algebraic multigrid solvers for physical systems¹² we apply a multiscale approach for recursively reducing the normalized-cut minimization problem in a non-iterative manner. We start by choosing about half of the pixels as representatives, which we call seeds: these are chosen so that every pixel in the original image is strongly coupled (that is, similar) to at least one seed adjacent to it⁶ (Fig. 2c). We then define the minimization problem only for the seeds, and subsequently approximate the solution for the whole image using an interpolation matrix that is set according to the coupling between neighbouring pixels in the fine scale (see Methods).

In addition to significantly reducing the number of nodes in the graph, this coarsening creates small aggregates of pixels adapted to the image at hand, the intensities of which are similar. Every pixel belongs to either one or several aggregates, each centred at one seed, by an interpolation weight proportional to the coupling between that pixel and the seed. We continue recursively (Fig. 2d), aggregating collections of nodes into much fewer nodes in the coarser-level graphs, thus creating a pyramid of graphs with larger aggregates of pixels in its coarser levels. Salient segments emerge in the appropriate level of the pyramid as nodes that are coupled weakly to their neighbours (for example, Figs 1b and 2d). Consequently, the minimization problem is simplified into one of looking for salient nodes at all levels of the pyramid.

Moreover, we use this approach to go far beyond cut minimization. After each coarsening step, coarse measurements are taken over the newly formed aggregates and are subsequently used to affect the aggregation process. (One such measure, shown in Supplementary Fig. 1, is the average intensity of the aggregate.) These multiscale measures for each aggregate include statistics characterizing the textures that appear in the image: the variance of the mean intensities and the second-order shape moments of its sub-aggregates at each finer level¹³. Consequently, each aggregate is represented by a multiscale set of characteristic measurements, efficiently summarizing its detailed pixel information. These measures are calculated recursively, and hence very efficiently: every measure at any level is determined directly from measures computed at the previous, finer levels. This treatment of texture replaces the need to apply many filters to the image^{4,14,15}, which suffers from the chicken-and-egg difficulty and from higher complexity.

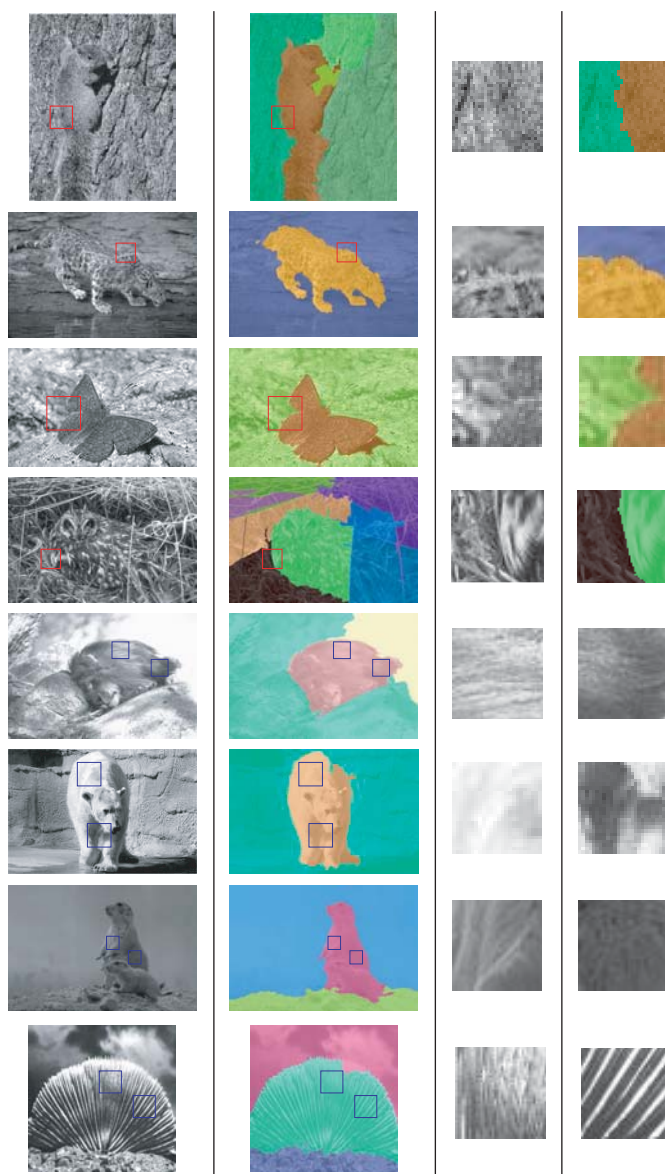


Figure 3 | Segmentation results for eight challenging images of animals on cluttered backgrounds. Transparent colours are overlaid over the original image (left) to mark segments (second left). In the top four images, we zoom into regions (marked with red squares) that are strikingly ambiguous to segment, and show our successful global-segmentation results. In the lower four images we zoom into regions (marked with blue squares) that have very different textural properties, but were nevertheless correctly grouped together into the same segment.

All of these measurements are incorporated into the segmentation process by affecting the coupling (similarity) between neighbouring aggregates (see Methods). Thus, analogous to the coupling between pixels according to similarity in luminance level, the coupling between aggregates reflects statistical similarities that could not be detected at finer levels. Note that this means that we are using contrast between region properties as a qualitative improvement of the process, in addition to contrast across boundaries in the simple normalized-cut approach.

The coarsening process detailed above keeps—out of the vast amount of data in the image—only the information necessary to segment it. In particular, nodes at coarse levels do not carry with them the precise boundary location of the corresponding segment. A top-down process can be used upon the detection of any salient segment in order to locate its exact boundary. We roll down the



Figure 4 | Similarity search by parts. Top: three example queries (cyan, pink, orange frames) with the original images on the left. The segmentation of each image is shown on the right in each frame, with the target query segment (user-selected) outlined in yellow: lower flank of left frame, right

lens and left lens for the left, middle and right examples, respectively.

Bottom: 27 images from the database (out of 110; see Supplementary Fig. 3), with coloured boxes indicating the most similar sunglasses found for each query.

pyramid to a desired fine level, by multiplying the interpolation matrices relating each two consecutive levels. We take advantage of this top-down procedure to improve further the segmentation by an additional energy-minimization sweep at each level as it is encountered (see Methods), together with prodding the weighted interpolations towards boolean associations, thus sharply delineating the segment boundary.

The algorithm is extremely efficient, as only its initial, very simple aggregation is done at the level of individual pixels. Complexity per aggregate increases linearly with the level, whereas their number drops geometrically. Moreover, the algorithm can use massive parallel processing, especially at the finer (the most expensive) levels. Our current implementation takes 2 s to complete the bottom-up aggregation of a 450×450 image using a Pentium 4, 1.6-GHz processor. Further optimization is still possible for the runtime to take significantly less than a second, and even down to a tiny fraction of a second on a parallel computer.

We tested our method on a set of challenging natural images containing animals camouflaged against their backgrounds. Whereas humans may use object knowledge aided by memory to segment such images, our approach has been to find out how far segmentation can reach using input-driven processing only. Figure 3 shows a typical set of results. In each case we present the automatically detected level in the pyramid containing the most salient segment. In all examples the animal was segmented in one piece by our method, out-performing other leading algorithms (see Supplementary Fig. 2). Our method succeeds in these challenging images, in which a cluttered background is often locally difficult to distinguish from the animal segment. In the top four examples we zoom in on an area where local differences in the original image (second panel from right) are hard to detect, yet the multiscale considerations in our algorithm yield a successful segmentation (far right panel). Note, for example, the fine difference detected along the squirrel's back. Furthermore, despite differences between areas within a segment, our method captures the essential similarities well enough to join them together into a coherent segment. This is exemplified in the bottom four examples, which show two widely differing areas that are segmented together (two right panels). Two regions with very different luminance levels are segmented together in the lion example, because of texture similarity. Even such seemingly different areas as those belonging to the shell (bottom example) are joined together because their orderly oriented texture makes them more similar to each other than to neighbouring areas. In Supplementary Fig. 2 we compare SWA over the same set of images to several available state-of-the-art segmentation algorithms, and demonstrate its superior performance.

To demonstrate the utility of the hierarchical segment representation

of the image provided by the SWA for higher-level visual tasks, we next applied it to a search task. The task was to find within a database of objects—sunglasses in this case—those that most resemble a target item in some feature (for example, shape of lenses), by matching aggregated properties of salient segments in the SWA hierarchy. The system pre-segments all images, and a query consists of selecting a segment within the target image. The segment with the most similar aggregated properties (shape, colour) within all database images is instantaneously found by searching through all summarized properties, and the sunglasses it belongs to are presented as the most similar to the query image. Figure 4 presents the results of three queries using this system, and Supplementary Fig. 3 presents the full database. The success of this system highlights the value of a robust hierarchical segmentation for higher-level tasks, importantly allowing a comparison of the same semantic object parts (for example, lenses) within different images, thus comparing 'apples to apples' and 'oranges to oranges'.

Our SWA algorithm constructs a representation of the image as a hierarchy of adaptive segments and finds the most salient segments without predetermining their number or size. Many common objects can naturally be described as a hierarchical collection of segments—thus, obtaining a hierarchy of salient segments is a useful novel framework within which to tackle object recognition. Our framework naturally allows the incorporation of top-down effects, expressing prior knowledge about properties of visual objects, as well as effects of context and attention, through the modification of couplings in the adaptive structure. For example, a probabilistic preference for smooth edges can be used in a top-down manner to join together aggregates with co-aligned boundaries even when lighting conditions make them appear different¹⁶. The simple top-down process already implemented in our algorithm suggests that although bottom-up mechanisms quickly segment the image into meaningful regions, feedback is needed for the accurate delineation of segment boundaries, as suggested also for the human visual system¹⁷. The increase in the size of aggregates and the complexity of their characteristics when moving up the hierarchy resemble well known properties of the primate visual system, as does the importance of interactions between the different levels¹. Precise localization of segment borders may well be an important role for the massive feedback projections in the visual cortex.

METHODS

The image is regarded as a weighted graph $G = (V, E)$, V being its set of n nodes, each corresponding to a pixel, v_i ($i = 1, \dots, n$), and E the set of undirected weighted edges w_{ij} , connecting neighbouring nodes v_i, v_j . $w_{ij} = e^{-\alpha|I_i - I_j|}$, where I_i, I_j are the intensities of pixels i, j , respectively, and α is a globally set, positive real constant. We conveniently treat W as a symmetric matrix, with all $w_{ii} = 0$

and $w_{ij} = 0$ if v_i and v_j are not neighbours. We further associate with the nodes a state vector $\mathbf{u} = (u_1, u_2, \dots, u_n)$, and define the energy

$$\Gamma(\mathbf{u}) = \frac{\sum_{i>j} w_{ij}(u_i - u_j)^2}{\sum_{i>j} w_{ij}u_iu_j} = \frac{\mathbf{u}^T L \mathbf{u}}{\frac{1}{2} \mathbf{u}^T W \mathbf{u}} \quad (1)$$

where L , the so-called laplacian matrix of the graph, is set to satisfy the equality between the numerators, and W between the denominators. Any boolean assignment of $\mathbf{u}$ that yields a low-energy value $\Gamma(\mathbf{u})$ corresponds to a salient segment S in the image: the pixels $i \in \{1, 2, \dots, n\}$ for which $u_i = 1$ are the pixels in S , otherwise $u_i = 0$.

If we permit real-value assignments to $\mathbf{u}$, the minimum for Γ is obtained by the solution of the generalized eigen problem $L\mathbf{u} = \lambda W\mathbf{u}$ with minimal positive eigenvalue λ . The algebraic multigrid procedure for solving this eigen problem consists of choosing a representative subset of $N \approx \frac{n}{2}$ pixels in the image, which we call seeds and denote by renaming their corresponding state variables u_s to be $\mathbf{U} = (U_1, U_2, \dots, U_N)$. For suitable choice of representatives⁶, the minimizing eigenvector satisfies $\mathbf{u} = P\mathbf{U}$, where the interpolation P is a sparse matrix $\{p_{ij}\}$; assuming for simplicity of notation that $U_k = u_{k_s}$ ($k = 1, 2, \dots, N$), then $p_{ik} = w_{ik}/\sum_j w_{ij}$ for $i > N$; whereas for $i \leq N$, $p_{ik} = 0$ except for $p_{ii} = 1$, ($1 \leq i, k \leq N$). Substituting the interpolation relation in equation (1), we seek to minimize

$$\Gamma(\mathbf{U}) \approx \frac{\mathbf{U}^T (P^T L P) \mathbf{U}}{\frac{1}{2} \mathbf{U}^T (P^T W P) \mathbf{U}} = \frac{\sum_{k>l} \tilde{w}_{kl}(U_k - U_l)^2}{\sum_{k>l} \tilde{w}_{kl} U_k U_l} \quad (2)$$

where $\{\tilde{w}_{kl}\}$ are set to satisfy the equality between the numerators, and $\{\hat{w}_{kl}\}$ between the denominators. $\Gamma(\mathbf{U})$ may effectively be approximated by replacing $\{\tilde{w}_{kl}\}$ with the simpler $\{\hat{w}_{kl}\}$. We call $\{\hat{w}_{kl}\}$ the coarse graph weights. By this coarsening we reduce the original minimization problem in $\mathbf{u}$ to a much smaller minimization problem in $\mathbf{U}$, the solution of which approximates the solution in $\mathbf{u}$ via the interpolation relation. Finally, this coarsening procedure can be repeated recursively, level after level⁷.

We next modify the coarse graph weights to reflect also contrast in properties at the current coarse level, in addition to contrast at the finer level. This is done by collecting a vector of multilevel properties $\mathbf{f}_k = (f_{k1}, f_{k2}, \dots, f_{km})$ for every coarse node k and using these properties to reduce the coupling between neighbouring aggregates k and l by a factor proportional to $\exp(-\mathbf{f}_k^T \Lambda \mathbf{f}_l)$, where the diagonal matrix Λ weighs the importance of every property. Entries in $\mathbf{f}_k$ represent statistics over the properties of the set of sub-aggregates of k . These are computed from the sub-aggregates using averaging weighted according to the interpolation matrix, and include its average intensity and the variances in the average intensities of its sub-aggregates at all finer scales, as well as its low-order shape moments (based on averaging $x_l^i y_l^j$, where (x_l, y_l) is the location of node v_l and $k, l = 0, 1, \dots$).

We detect the salient segments at all levels of the pyramid as those aggregates k for which $\Gamma(\mathbf{U})$ has low values, with the state vector $\mathbf{U}$ set to 1 at the k th entry and 0 elsewhere. For each such aggregate k we use a top-down process to delineate its boundary. We start at the level at which k was detected with its characteristic state vector $\mathbf{U}$. By repeating interpolations from this level down to any finer level, using successively the interpolation relations given by the matrices P , we obtain for each finer aggregate the relative weight by which it relates to the aggregate k . Our goal at this stage is to bring $\mathbf{u}$ closer to a boolean state vector. We do so by modifying, at each of the finer levels, the interpolated $\mathbf{u}$ before interpolating it to the next finer level. Values $u_i > 0.9$ are set to 1, and values $u_i < 0.1$ are set to 0. Finally, at the finest level each pixel is associated solely with the aggregate k at the coarsest level for which its weight turned out largest. Note that despite the

incorporation of the top-down process, complexity remains linear because the finest scale of aggregates needed to detail the segment boundary is proportional to its size.

Received 25 March; accepted 13 June 2006.
Published online 28 June 2006.

1. Felleman, D. J. & Van Essen, D. C. Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* **1**, 1–47 (1991).
2. Pietikainen, M., Rosenfeld, A. & Walter, I. Split-and-link algorithms for image segmentation. *Pattern. Recogn.* **15**, 287–298 (1982).
3. Comaniciu, D. & Meer, P. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern. Anal. Machine Intell.* **24**, 603–619 (2002).
4. Malik, J., Belongie, S., Leung, T. & Shi, J. Contour and texture analysis for image segmentation. *Int. J. Comput. Vision* **43**, 7–27 (2001).
5. Felzenszwalb, P. & Huttenlocher, D. Efficient graph-based image segmentation. *Int. J. Comput. Vision* **59**, 167–181 (2004).
6. Brandt, A. Algebraic multigrid theory: the symmetric case. *Appl. Math. Comput.* **19**, 23–56 (1986).
7. Sharon, E., Brandt, A. & Basri, R. Fast multiscale image segmentation. *Proc. IEEE Conf. Comput. Vision Pattern Recogn.* **1**, 70–77 (2000).
8. Hubel, D. H. & Wiesel, T. N. Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* **195**, 215–243 (1968).
9. Tanaka, K. Inferotemporal cortex and object vision. *Annu. Rev. Neurosci.* **19**, 109–139 (1996).
10. Shi, J. & Malik, J. Normalized cuts and image segmentation. *IEEE Trans. Pattern. Anal. Machine Intell.* **22**, 888–905 (2000).
11. Golub, G. H. & Van Loan, C. F. *Matrix Computations* (Johns Hopkins Univ. Press, Baltimore, 1989).
12. Brandt, A., McCormick, S. & Ruge, J. In *Sparsity and its Applications* (ed. Evans, D. J.) 257–284 (Cambridge Univ. Press, Cambridge, 1984).
13. Galun, M., Sharon, E., Basri, R. & Brandt, A. Texture segmentation by multiscale aggregation of filter responses and shape elements. *Proc. Int. Conf. Comput. Vision* **1**, 469–476 (2003).
14. Julesz, B. Textons, the elements of texture perception, and their interactions. *Nature* **290**, 91–97 (1981).
15. Voorhees, H. & Poggio, T. Computing texture boundaries from images. *Nature* **333**, 364–367 (1988).
16. Sharon, E., Brandt, A. & Basri, R. Segmentation and boundary detection using multiscale intensity measurements. *Proc. IEEE Conf. Comput. Vision Pattern Recogn.* **1**, 469–476 (2001).
17. Stanley, D. A. & Rubin, N. fMRI activation in response to illusory contours and salient regions in the human lateral occipital complex. *Neuron* **37**, 323–331 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements Research was supported in part by the European Commission Project Aim Shape, the Binational Science foundation, and by the German-Israeli Foundation. D.S. was supported by a grant from the National Institutes of Health. The research was conducted at the Moross Laboratory for Vision and Motor Control at the Weizmann Institute of Science. We thank N. Rubin and D. Jacobs for many useful remarks, and S. Geman for commenting on an earlier version of the manuscript. We are grateful to E. Borenstein for his help with constructing the sunglasses search system. We also thank M. Varma and R. Deitch for help with the comparisons presented in the Supplementary Information and N. Brandt for help with the graphics.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to E.S. (aitan.sharon@weizmann.ac.il).

LETTERS

Phosphorylation of WAVE1 regulates actin polymerization and dendritic spine morphology

Yong Kim¹, Jee Young Sung¹, Ilaria Ceglia¹, Ko-Woon Lee¹, Jung-Hyuck Ahn¹, Jonathan M. Halford¹, Amie M. Kim¹, Seung P. Kwak², Jong Bae Park³, Sung Ho Ryu³, Annette Schenck⁴, Barbara Bardoni⁴, John D. Scott⁵, Angus C. Nairn^{1,6} & Paul Greengard¹

WAVE1—the Wiskott–Aldrich syndrome protein (WASP)-family verprolin homologous protein 1—is a key regulator of actin-dependent morphological processes¹ in mammals, through its ability to activate the actin-related protein (Arp2/3) complex. Here we show that WAVE1 is phosphorylated at multiple sites by cyclin-dependent kinase 5 (Cdk5) both *in vitro* and in intact mouse neurons. Phosphorylation of WAVE1 by Cdk5 inhibits its ability to regulate Arp2/3 complex-dependent actin polymerization. Loss of WAVE1 function *in vivo* or in cultured neurons results in a decrease in mature dendritic spines. Expression of a dephosphorylation-mimic mutant of WAVE1 reverses this loss of WAVE1 function in spine morphology, but expression of a phosphorylation-mimic mutant does not. Cyclic AMP (cAMP) signaling reduces phosphorylation of the Cdk5 sites in WAVE1, and increases spine density in a WAVE1-dependent manner. Our data suggest that phosphorylation/dephosphorylation of WAVE1 in neurons has an important role in the formation of the filamentous actin cytoskeleton, and thus in the regulation of dendritic spine morphology.

Cdk5, together with its neuron-specific regulatory subunit P35, has been implicated in neurite outgrowth and neuronal migration during brain development^{2,3}, in striatal function in mature brain⁴, and in Alzheimer's disease^{3,5} and drug addiction^{6,7}. In view of the growing appreciation of Cdk5 in normal and abnormal neuronal function, we undertook a search for interacting proteins. To identify interacting proteins, purified recombinant P35/Cdk5 protein was immobilized on Protein A beads coupled to anti-P35 antibody, and affinity chromatography was used to isolate proteins from striatal lysates (Fig. 1a). Three major co-precipitated proteins with a relative molecular mass (M_r) of 140,000, 125,000 and 43,000 were identified as PIR121 (p53-inducible messenger RNA), Nap1 (Nck-associated protein 1) and β -actin respectively, by matrix-assisted laser desorption/ionization–time of flight (MALDI-TOF) mass spectrometry (data not shown). Previously, PIR121 and Nap1 were co-purified from bovine brain with WAVE1, Abi2 (Abl interactor 2) and HSPC300 in a pentameric protein complex⁸. The presence of WAVE1 (but not WAVE2 or WAVE3), PIR121 and Nap1 in the complex of P35/Cdk5 binding proteins was confirmed by immunoblotting with specific antibodies (Fig. 1b, Supplementary Fig. 1a). Evidence for the interaction of the WAVE1 complex and P35 was also obtained from co-immunoprecipitation studies: (1) using lysates prepared from cultured neurons (Supplementary Fig. 1b), and (2) after expression of proteins in COS-7 cells (Supplementary Fig. 1c). Direct interaction of P35 (but not Cdk5) with PIR121, Nap1 and WAVE1 was also found using a 'far western' blotting technique (Supplementary

Fig. 1d–f), suggesting a critical role for P35 in the interaction with the WAVE1 complex.

Purified WAVE1 complex was incubated without or with purified P35/Cdk5 in the presence of Mg[γ -³²P]ATP. In addition to P35 (which is known to be autophosphorylated by Cdk5), WAVE1 was highly phosphorylated, with an upward mobility shift (Fig. 1c). Cdk5 is a so-called proline-directed kinase that phosphorylates serine or threonine residues with an adjacent carboxy-terminal proline residue. Three serines, S310 as the major site (~50% of phosphorylation) and S397 and S441 as minor sites, were identified (Supplementary Fig. 2a). Triple mutations at S310, S397 and S441 abolished ~83% of phosphorylation (data not shown). The three phosphorylation sites are located in the proline-rich region of WAVE1 (Fig. 1d), but are not conserved in WAVE2 or WAVE3 (ref. 9). Phosphorylation by Cdk5 may therefore be specific for WAVE1. However, it is possible that WAVE2 and WAVE3 are phosphorylated at other non-conserved proline-directed sites by Cdk5 or other kinases. We also confirmed the phosphorylation at S310, S397 and S441 using phospho-specific antibodies (Supplementary Fig. 2b, c).

High basal phosphorylation of S310, S397 and S441 was detected in striatal slices (see Supplementary Table 1). Treatment with the Cdk5 inhibitor, roscovitine, resulted in a marked reduction of phosphorylation at each of the three sites (S310, 23%; S397, 11%; S441, 40% of control level with 10 μ M roscovitine) (Fig. 1e). Phosphorylation of WAVE1 was not affected by treatment with U0126, an inhibitor of mitogen-activated protein kinase kinase (MAPKK) (Fig. 1e). Phosphorylation of MAPK (another proline-directed kinase) was reduced to near undetectable levels after treatment with U0126, but not with roscovitine. Moreover, analysis of primary cultured cortical neurons from Cdk5-knockout mice revealed a significant reduction of phosphorylation at each of the three sites (S310, 44 $\pm$ 11.9%; S397, 18 $\pm$ 4.1%; S441, 42 $\pm$ 9.7% of wild-type level) (Supplementary Fig. 3). These results indicate that WAVE1 is a physiological substrate for Cdk5, but it also appears that additional kinase(s) phosphorylate WAVE1 in neurons.

WAVE and related WASP proteins have a conserved verprolin-homology, cofilin-homology, acidic (VCA) domain that directly binds to and activates the Arp2/3 complex, resulting in polymerization of actin¹⁰. The WAVE1 complex purified from rat brain cytosol and recombinant His-WAVE1 were both highly active, and comparable to a GST-VCA domain from WASP in the pyrene actin assay (Supplementary Fig. 4). The high activity of the WAVE1 complex was dependent on WAVE, but not on contaminating WASP proteins. N-WASP was not present in the WAVE1 complexes, as judged by immunoblotting (data not shown). Moreover, pre-incubation of the

¹Laboratory of Molecular and Cellular Neuroscience, The Rockefeller University, 1230 York Avenue, New York, New York 10021, USA. ²Discovery Neurosciences, Wyeth Research, Princeton, New Jersey 08543, USA. ³Division of Molecular and Life Sciences, Pohang University of Science and Technology, Pohang, 790-784, Korea. ⁴Institut de Génétique et de Biologie Moléculaire et Cellulaire, CNRS/INSERM/ULP, BP 10142, 67404 Illkirch cedex, France. ⁵Howard Hughes Medical Institute, Vollum Institute, 3181 Sam Jackson Park Road, Portland, Oregon 97239-3098, USA. ⁶Department of Psychiatry, Yale University School of Medicine, 34 Park Street, New Haven, Connecticut 06508, USA.

WAVE1 complex (or recombinant WAVE1) with an anti-C-terminal WAVE1 antibody abolished Arp2/3-complex-dependent actin polymerization, while having no effect on the activity of the GST-VCA domain from WASP (data not shown). WAVE2 complex reconstituted *in vitro* was also found to be active¹¹, supporting our conclusion that the WAVE1 complex is active. Our results contrast with a previous result indicating that the WAVE1 complex purified from bovine brain was inactive⁸.

We next examined the effect of phosphorylation by Cdk5 on WAVE1-induced Arp2/3-complex-mediated actin polymerization. Pre-incubation of the WAVE1 complex (Fig. 2a) or of recombinant WAVE1 (Fig. 2b) with P35/Cdk5 in the presence of MgATP led to a substantial reduction in actin polymerization compared to that obtained when MgATP was omitted. Notably, basal phosphorylation at S310, S397 and S441 was found in recombinant WAVE1 (insets in Fig. 2b, c, see also Supplementary Table 1). Pre-incubation of recombinant WAVE1 with a non-specific λ protein phosphatase (λ PPase) reduced basal phosphorylation and this was accompanied by an increase in WAVE1-mediated actin polymerization (Fig. 2c).

Mutation of S310 (S310A), S397 (S397A), or S441 (S441A) to alanine rendered WAVE1 activity partially resistant to inhibition by Cdk5-dependent phosphorylation: the generation of barbed ends by WAVE1 after phosphorylation was 29%, 72%, 50% and 41% for the wild-type, S310A, S397A, and S441A mutants, respectively, compared to the corresponding unphosphorylated proteins. Combined mutation at S310 and S397 was no more effective than mutation at S310 alone (Fig. 2d and Supplementary Fig. 5b, c). Conversely, λ PPase-mediated activation was not observed following

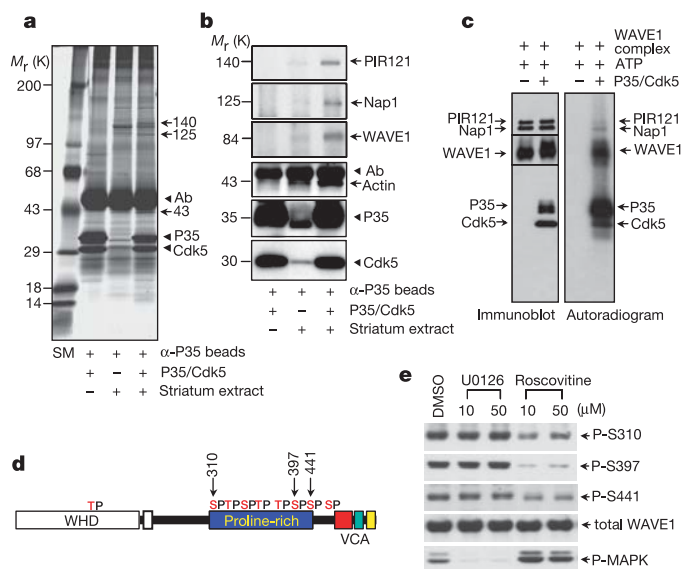


Figure 1 | WAVE1 interacts with, and is phosphorylated by, P35/Cdk5.

a, b, Purified P35/Cdk5 was immobilized on anti-P35 antibody-coupled beads, without or with striatal extract. Bound proteins were visualized with silver stain (**a**) or immunoblotted for the indicated proteins (**b**). Arrowheads indicate immobilized antibody (Ab), P35 or Cdk5. **c**, Purified WAVE1 complex was incubated with [γ -³²P]ATP in the absence or presence of P35/Cdk5, and the samples were analysed, following SDS-polyacrylamide gel electrophoresis (SDS-PAGE), by immunoblotting for the indicated proteins, or by autoradiography. **d**, Domain structure of WAVE1 including the proline-rich domain, amino-terminal WAVE homology domain (WHD) and C-terminal VCA domain. Nine TP or SP (threonine-proline or serine-proline) sites are found in both mouse and human WAVE1. Arrows indicate Cdk5 phosphorylation sites. **e**, Striatal slices were incubated for 1 h with DMSO, a Cdk5 inhibitor (roscovitine, 10 μ M, 50 μ M) or a MAPK inhibitor (U0126, 10 μ M, 50 μ M). Phosphorylation of WAVE1 at S310, S397 and S441, total WAVE1, and phospho-MAP kinase (P-MAPK) were analysed by immunoblotting.

mutation of S310 alone or double mutation of S310A and S397A (Fig. 2e and Supplementary Fig. 6). The S310A, S397A, S441A, and S310A:S397A mutants exhibited comparable basal activity to wild-type WAVE1 (Supplementary Fig. 5a). In contrast, a mutant form of WAVE1 where S310 was changed to aspartate (S310D) to mimic phosphorylation exhibited low basal activity that was little affected by incubation with P35/Cdk5 or λ PPase (Fig. 2d, e). Altogether, the results indicate that phosphorylation of S310 by Cdk5 is largely responsible for inhibition of WAVE1 activity, with phosphorylation of S397 playing a lesser role and phosphorylation of S441 having little influence.

WASP, as well as the WAVE-binding proteins Abi2 and insulin receptor tyrosine kinase substrate p53 (IRSp53), have been implicated in the regulation of spine morphology^{12–14}, and formation of branched actin filaments in spine heads¹⁵ probably involves the Arp2/3 complex. However, little is known about the role of WAVE1 in dendritic spine morphology. We next analysed spine morphology of medium spiny neurons in striatal sections obtained from WAVE1^{-/-} mice.

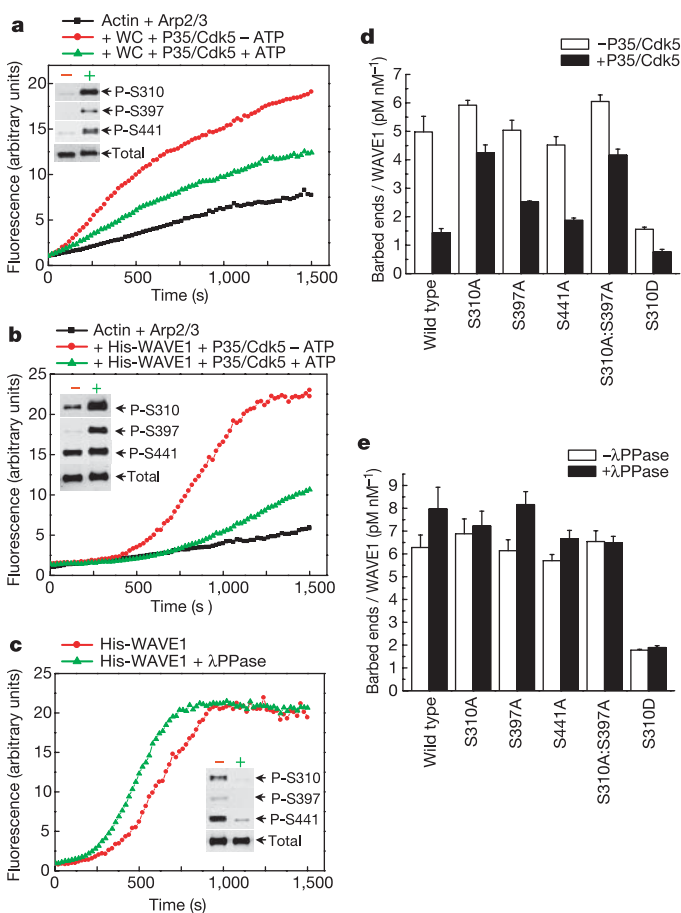


Figure 2 | Inhibition of Arp2/3-complex-mediated actin polymerization by phosphorylation of WAVE1.

a, b, Native WAVE1 complex (WC) (**a**) or recombinant His-WAVE1 (**b**) was pre-incubated with P35/Cdk5 in the absence or presence of MgATP, and a pyrene-actin polymerization assay performed. The lower inhibition of the WAVE1 complex compared to recombinant WAVE1 results from the lower fractional phosphorylation of WAVE1 at S310 in the WAVE1 complex (see Supplementary Table 1). **c**, His-WAVE1 was pre-incubated in the absence or presence of λ PPase and actin polymerization was measured. In **a–c**, the insets show phospho-WAVE1 and total WAVE1 in the absence (–) and presence (+) of MgATP or λ PPase. **d, e**, The activity of wild-type WAVE1 or phosphorylation site mutants was measured after (**d**) pre-incubation with MgATP in the absence or presence of P35/Cdk5, or (**e**) with or without λ PPase. The barbed-end generation by WAVE1 was calculated from data shown in Supplementary Figs 5b and 6a. Error bars indicate the standard error of the linear regression analysis of data obtained from duplicate actin polymerization assays.

WAVE1^{-/-} neurons displayed a significant reduction in class 1 spines (stubby type), class 2 spines (comprised of mature spines with a relatively large head and a short neck) and class 3 spines (thin spines) and showed a relatively high density of class 4 protrusions, or filopodia (Fig. 3a, b). Similarly, reduced expression of WAVE1 (as a result of RNAi) in primary hippocampal neurons resulted in a decrease in the density of class 2 spines and an increase in filopodia (Fig. 3c, d). Expression of RNAi-resistant wild-type (Rr-WT) WAVE1 (in conjunction with RNAi) reversed the loss of the class 2 spines as well as the increase in filopodia. Notably, expression of a Rr-S310A mutant (to mimic the dephosphorylated active form) had a greater effect than Rr-WT, whereas expression of a Rr-S310D mutant (to mimic the phosphorylated inactive form) had little or no effect on the RNAi-induced phenotype (Fig. 3e, f).

Our results indicate that WAVE1 is phosphorylated at three sites in intact neurons under basal conditions (Supplementary Table 1), reflecting basal activity of Cdk5. Stimulation of striatal slices with the dopamine D1 agonist SKF82526 decreased WAVE1 phosphorylation at S310, S397 and S441 (60 ± 9.1 , 29 ± 4.3 , $36 \pm 11.8\%$ of control, respectively) (Fig. 4a). Similar results were obtained using forskolin to activate the cAMP pathway (Fig. 4b). Activation of the

cAMP pathway had no effect on Cdk5 activity (Supplementary Fig. 7), suggesting the involvement of a protein phosphatase in cAMP-mediated WAVE1 dephosphorylation. When pure neuronal cultures of hippocampus were treated with forskolin for 1 h, the level of phosphorylation of all three sites decreased (Fig. 4c). (Similar results were obtained for cortical neurons; data not shown.) This was associated with an increase in the numbers of class 2 and 3 spines as well as class 4 protrusions (Fig. 4d, e). Downregulation of WAVE1 using RNAi abolished the forskolin-induced increase in the number of class 2 and 3 spines (Fig. 4d, e).

The present results indicate that WAVE1 phosphorylation/dephosphorylation plays an important part in the regulation of actin polymerization and dendritic spine morphology. The WAVE proteins

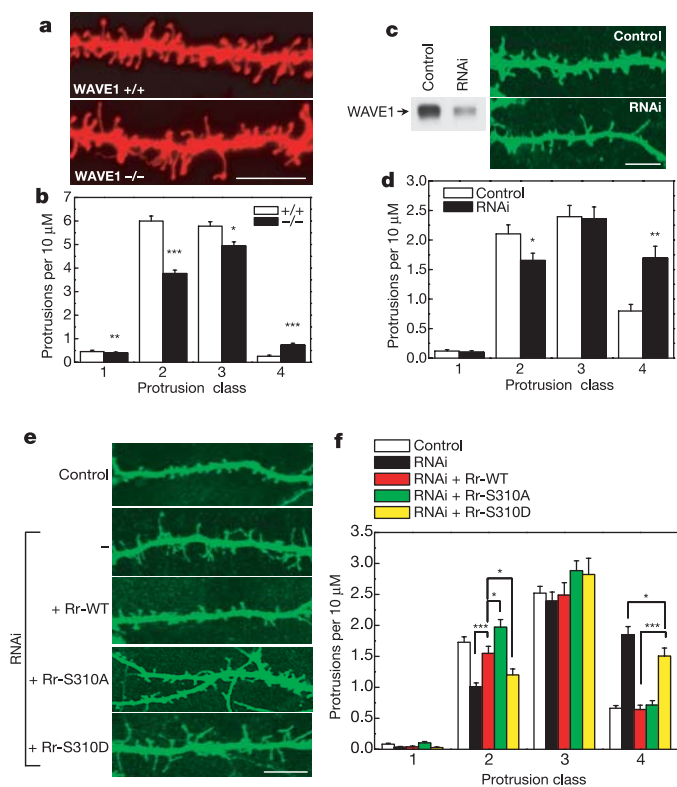


Figure 3 | Role of WAVE1 in spine morphogenesis. **a, b**, Spine morphology of medium spiny neurons in striatum of wild-type (+/+) or WAVE1 knockout (-/-) mice (P24) (visualization used DiI staining). **c, d**, Primary hippocampal neurons (8 days *in vitro*) were co-transfected with EGFP plus WAVE1-RNAi plasmid (RNAi) or a plasmid vector (Control). Spine morphology (12 days *in vitro*) was visualized in **c** and **e** using anti-GFP antibody. Downregulation of WAVE1 was confirmed by immunoblot analysis of neurons infected with a lentiviral vector containing WAVE1 RNAi (left in **c**). **e, f**, Neurons were co-transfected with EGFP plus control plasmid vectors (Control) or plasmids for RNAi and RNAi-resistant WAVE1, as indicated. Wild-type and mutant Rr-WAVE1 transgenes contained an N-terminal Myc-tag and equivalent co-expression of transgenes was observed upon immunostaining of Myc (data not shown). Scale bars in **a, c** and **e**, 10 μ m. **b, d, f**, Quantification of protrusion density (number per 10 μ m dendrite length; mean $\pm$ s.e.m.) in four classes. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$ versus wild type (**b**), Control (**d**), or RNAi or RNAi + Rr-WT as indicated (**f**), Kolmogorov-Smirnov test.

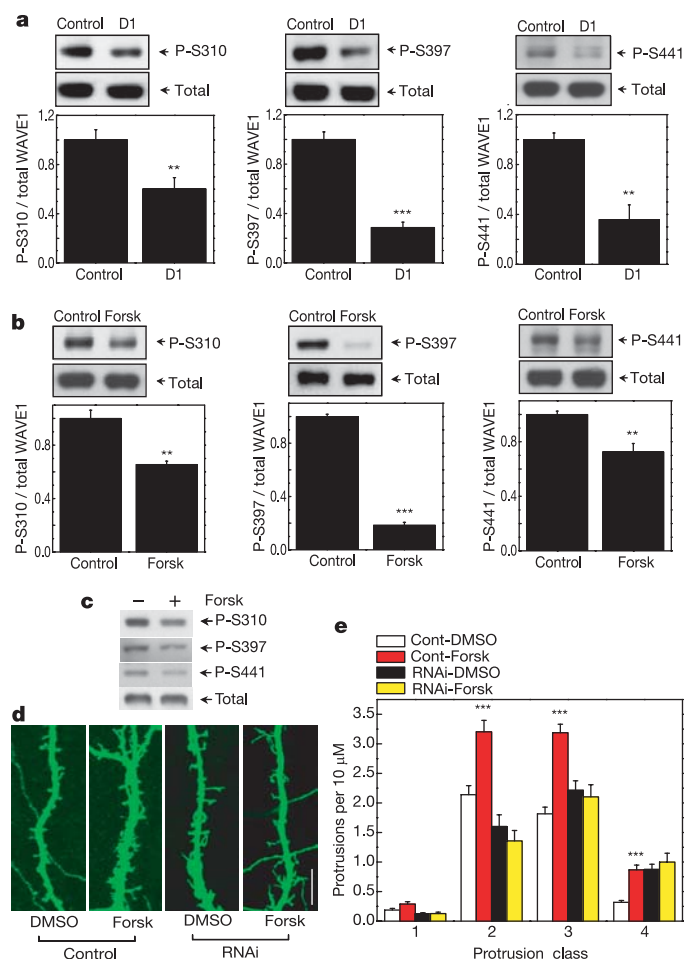


Figure 4 | cAMP-dependent dephosphorylation of WAVE1 and its role in spine formation. **a, b**, Striatal slices were incubated in the presence of DMSO (Control), the dopamine D1 receptor agonist SKF82526 (D1, 1 μ M) (**a**) or forskolin (Forsk, 1 μ M) (**b**) for 30 min. The levels of phospho- and total WAVE1 were measured by immunoblotting (upper panels) and the results quantified by densitometry (lower panels). Data, normalized to control values, represent means $\pm$ s.e.m. for four to six experiments. ** $P < 0.01$, *** $P < 0.001$ versus control, Student's *t* test. **c**, Primary hippocampal neurons were incubated with DMSO (-) or forskolin (+, 1 μ M) for 1 h, and phospho-WAVE1 and total WAVE1 visualized by immunoblotting. **d, e**, Primary hippocampal neurons were co-transfected with EGFP plus WAVE1-RNAi plasmid (RNAi) or a plasmid vector (Control), and incubated with DMSO or forskolin (Forsk, 1 μ M) for 1 h. Dendritic spine morphology was visualized by immunostaining with anti-green fluorescent protein (GFP) antibody. Scale bar, 10 μ m. **e**, Quantification of protrusion density (number per 10 μ m dendrite length; mean $\pm$ s.e.m.) in four classes. *** $P < 0.001$ for forskolin versus DMSO, Kolmogorov-Smirnov test.

are activated by Rac¹⁶, a Rho family GTPase, and WAVE1 may be involved in mediating the effects of Rac on spine formation^{17,18}. A previous study showed that WAVE1 knockout mice exhibit deficits in learning and memory¹⁹, a phenotype that may be related to the abnormal spine morphology characterized here. Spines contain both bundled and branched forms of filamentous actin¹⁵ and it is likely that WAVE1 contributes to the formation of spines, as well as to remodelling of the actin cytoskeleton in spines during development and in processes affecting synaptic plasticity²⁰. In mature neurons in brain or in culture, WAVE1 was highly phosphorylated at Cdk5 sites (S310 and S397 were phosphorylated to stoichiometries of ~0.7 and ~0.9 mol mol⁻¹, respectively, Supplementary Table 1). Stimulation of cAMP-dependent signalling resulted in dephosphorylation of the Cdk5 sites. Thus, WAVE1 would be largely in an inactive form under basal conditions, but would be activated by neurotransmitters, such as dopamine, that increase the levels of cAMP. WAVE1 was found to be required for the effect of cAMP on spine formation, and this probably occurs as a result of cAMP-dependent dephosphorylation and activation of WAVE1. Further clarification of the mechanisms by which Cdk5, cAMP and other signalling pathways regulate WAVE1 phosphorylation/dephosphorylation should lead to an increased understanding of the regulation of spine morphogenesis and synaptic function in developing and adult brain.

METHODS

Pull-down assay. Purified P35/Cdk5 (ref. 21) (20 µg) was incubated with anti-P35 antibody-coupled beads (15 µg antibody/60 µl agarose beads conjugated to protein A). The immobilized P35/Cdk5 was mixed with an extract from ten pooled rat striata homogenized in 20 mM Tris-HCl (pH 7.5), 2 mM MgCl₂, 150 mM NaCl, 1% Triton X-100, 30 mM NaF, 1 mM Na₃VO₄, 30 mM Na₄P₂O₇ and protease inhibitor cocktail (Roche). After overnight incubation, the beads were precipitated and washed five times with homogenization buffer. Co-precipitated proteins were separated by SDS-PAGE (6–16% acrylamide), followed by silver staining (BioRad) or immunoblotting.

Striatal slices. Striatal slices were prepared from male C57BL/6 mice (6–8 weeks old) as described in ref. 22.

Dil staining. Brain slices from wild-type or WAVE1^{-/-} mice²³ were prepared and labelled with lipophilic fluorescence dye 1,1'-diiododecyl-3,3',3'-tetramethylindocarbocyanine perchlorate (DiI) as described in ref. 24.

Actin polymerization assay. Pyrene-labelled and non-labelled actin (from rabbit muscle), and Arp2/3 complex were obtained from Cytoskeleton, Inc. WAVE1 complex or recombinant WAVE1 was mixed with Arp2/3 complex (60 nM) in 10 mM Tris (pH 7.5), 50 mM KCl, 1 mM MgCl₂, 1 mM EGTA and 0.5 mM ATP. After addition of pyrene-labelled actin (1 µM, 10% labelled), fluorescence was measured at 21 °C in a fluorimeter. The concentration of barbed ends was calculated²⁵ at the point where polymerization was 50% of the maximum measured within 1,500 s. The concentration of barbed ends measured in the absence of WAVE1 was subtracted from the values calculated in the presence of WAVE1, to obtain measurements of WAVE1-dependent actin polymerization. The concentration of barbed ends generated by wild-type and mutants of WAVE1 was plotted versus WAVE1 concentration, as shown in Supplementary Figs 4d, 5a, 5b and 6a. The data related to the generation of barbed ends by WAVE1 (shown in Fig. 2d and e) was derived from the initial gradient of the slopes shown in Supplementary Figs 5b and 6a.

Image analysis. Acquisition of fluorescence images, morphometric measurements and classification of dendritic spines were as described²⁴.

Additional information related to DNA constructs, phosphorylation and dephosphorylation of WAVE1 for the actin polymerization assay, cell culture and transfection, and analysis of spines, is included in the Supplementary Methods.

Received 18 April; accepted 6 June 2006.

Published online 16 July 2006.

- Stradal, T. E. *et al.* Regulation of actin dynamics by WASP and WAVE family proteins. *Trends Cell Biol.* **14**, 303–311 (2004).
- Smith, D. S., Greer, P. L. & Tsai, L. H. Cdk5 on the brain. *Cell Growth Differ.* **12**, 277–283 (2001).
- Cruz, J. C. & Tsai, L. H. A Jekyll and Hyde kinase: roles for Cdk5 in brain

- development and disease. *Curr. Opin. Neurobiol.* **14**, 390–394 (2004).
- Bibb, J. A. *et al.* Phosphorylation of DARPP-32 by Cdk5 modulates dopamine signalling in neurons. *Nature* **402**, 669–671 (1999).
- Patrick, G. N. *et al.* Conversion of p35 to p25 deregulates Cdk5 activity and promotes neurodegeneration. *Nature* **402**, 615–622 (1999).
- Bibb, J. A. *et al.* Effects of chronic exposure to cocaine are regulated by the neuronal protein Cdk5. *Nature* **410**, 376–380 (2001).
- Norrholm, S. D. *et al.* Cocaine-induced proliferation of dendritic spines in nucleus accumbens is dependent on the activity of cyclin-dependent kinase-5. *Neuroscience* **116**, 19–22 (2003).
- Eden, S., Rohatgi, R., Podtelejnikov, A. V., Mann, M. & Kirschner, M. W. Mechanism of regulation of WAVE1-induced actin nucleation by Rac1 and Nck. *Nature* **418**, 790–793 (2002).
- Suetsugu, S., Miki, H. & Takenawa, T. Identification of two human WAVE/SCAR homologues as general actin regulatory molecules which associate with the Arp2/3 complex. *Biochem. Biophys. Res. Commun.* **260**, 296–302 (1999).
- Takenawa, T. & Miki, H. WASP and WAVE family proteins: key molecules for rapid rearrangement of cortical actin filaments and cell movement. *J. Cell Sci.* **114**, 1801–1809 (2001).
- Innocenti, M. *et al.* Abi1 is essential for the formation and activation of a WAVE2 signalling complex. *Nature Cell Biol.* **6**, 319–327 (2004).
- Irie, F. & Yamaguchi, Y. EphB receptors regulate dendritic spine development via intersectin, Cdc42 and N-WASP. *Nature Neurosci.* **5**, 1117–1118 (2002).
- Grove, M. *et al.* ABI2-deficient mice exhibit defective cell migration, aberrant dendritic spine morphogenesis, and deficits in learning and memory. *Mol. Cell Biol.* **24**, 10905–10922 (2004).
- Choi, J. *et al.* Regulation of dendritic spine morphogenesis by insulin receptor substrate 53, a downstream effector of Rac1 and Cdc42 small GTPases. *J. Neurosci.* **25**, 869–879 (2005).
- Rao, A. & Craig, A. M. Signaling between the actin cytoskeleton and the postsynaptic density of dendritic spines. *Hippocampus* **10**, 527–541 (2000).
- Miki, H., Suetsugu, S. & Takenawa, T. WAVE, a novel WASP-family protein involved in actin reorganization induced by Rac. *EMBO J.* **17**, 6932–6941 (1998).
- Luo, L. *et al.* Differential effects of the Rac GTPase on Purkinje cell axons and dendritic trunks and spines. *Nature* **379**, 837–840 (1996).
- Nakayama, A. Y., Harms, M. B. & Luo, L. Small GTPases Rac and Rho in the maintenance of dendritic spines and branches in hippocampal pyramidal neurons. *J. Neurosci.* **20**, 5329–5338 (2000).
- Soderling, S. H. *et al.* Loss of WAVE-1 causes sensorimotor retardation and reduced learning and memory in mice. *Proc. Natl Acad. Sci. USA* **100**, 1723–1728 (2003).
- Matus, A. Actin-based plasticity in dendritic spines. *Science* **290**, 754–758 (2000).
- Saito, T. *et al.* Developmental regulation of the proteolysis of the p35 cyclin-dependent kinase 5 activator by phosphorylation. *J. Neurosci.* **23**, 1189–1197 (2003).
- Nishi, A., Snyder, G. L. & Greengard, P. Bidirectional regulation of DARPP-32 phosphorylation by dopamine. *J. Neurosci.* **17**, 8147–8155 (1997).
- Dahl, J. P. *et al.* Characterization of the WAVE1 knock-out mouse: implications for CNS development. *J. Neurosci.* **23**, 3343–3352 (2003).
- Lee, K.-W. *et al.* Cocaine-induced dendritic spine formation in D1 and D2 dopamine receptor-containing medium spiny neurons in nucleus accumbens. *Proc. Natl Acad. Sci. USA* **103**, 3399–3404 (2006).
- Machesky, L. M. *et al.* Scar, a WASP-related protein, activates nucleation of actin filaments by the Arp2/3 complex. *Proc. Natl Acad. Sci. USA* **96**, 3739–3744 (1999).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank W.-B. Gan for assisting us with the Dil staining method; L.-H. Tsai for providing P35 cDNA; A. Yamamoto for Nap1 cDNA; and P. Aspenström for N-WASP antibody. We also thank T. D. Pollard for suggestions. This work was supported by the postdoctoral fellowship programme of the Korea Science & Engineering Foundation (KOSEF) (to J.Y.S.), and funding from the F.M. Kirby foundation (to P.G.), the Picower Foundation (to P.G.), the National Institute of Mental Health, the National Institute of Drug Abuse and the National Institute on Aging (to A.C.N. and P.G.).

Author Contributions Y.K., J.Y.S., I.C. and K.-W.L. performed experiments; J.M.H. and A.M.K. assisted in experiments; J.-H.A., S.P.K., A.S., B.B. and J.D.S. provided reagents; J.B.P. and S.H.R. performed MALDI-TOF mass spectrometry; Y.K., A.C.N. and P.G. designed experiments and wrote the manuscript.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to Y.K. (kimyo@rockefeller.edu), A.C.N. (angus.nairn@yale.edu) and P.G. (greengard@rockefeller.edu).

LETTERS

Transformation from committed progenitor to leukaemia stem cell initiated by MLL-AF9

Andrei V. Krivtsov¹, David Twomey^{1,5}, Zhaohui Feng², Matthew C. Stubbs¹, Yingzi Wang¹, Joerg Faber¹, Jason E. Levine^{1,2}, Jing Wang², William C. Hahn^{3,5}, D. Gary Gilliland^{4,6}, Todd R. Golub^{2,5,6} & Scott A. Armstrong^{1,2}

Leukaemias and other cancers possess a rare population of cells capable of the limitless self-renewal necessary for cancer initiation and maintenance^{1–7}. Eradication of these cancer stem cells is probably a critical part of any successful anti-cancer therapy, and may explain why conventional cancer therapies are often effective in reducing tumour burden, but are only rarely curative. Given that both normal and cancer stem cells are capable of self-renewal, the extent to which cancer stem cells resemble normal tissue stem cells is a critical issue if targeted therapies are to be developed. However, it remains unclear whether cancer stem cells must be phenotypically similar to normal tissue stem cells or whether they can retain the identity of committed progenitors. Here we show that leukaemia stem cells (LSC) can maintain the global identity of the progenitor from which they arose while activating a limited stem-cell- or self-renewal-associated programme. We isolated LSC from leukaemias initiated in committed granulocyte macrophage progenitors through introduction of the MLL–AF9 fusion protein encoded by the t(9;11)(p22;q23). The LSC were capable of transferring leukaemia to secondary recipient mice when only four cells were transferred, and possessed an immunophenotype and global gene expression profile very similar to that of normal granulocyte macrophage progenitors. However, a subset of genes highly expressed in normal haematopoietic stem cells was re-activated in LSC. LSC can thus be generated from committed progenitors without widespread reprogramming of gene expression, and a leukaemia self-renewal-associated signature is activated in the process. Our findings define progression from normal progenitor to cancer stem cell, and suggest that targeting a self-renewal programme expressed in an abnormal context may be possible.

Some fusion proteins encoded by translocations found in human leukaemias including those involving the mixed lineage leukaemia (MLL) gene have been reported to impart leukaemia stem cell properties on committed haematopoietic progenitors^{8–10}. We expressed an MLL–AF9 fusion protein in highly purified interleukin (IL)-7R⁺Lin[–]Sca-1[–]c-Kit⁺CD34⁺FcγRII/III^{hi} granulocyte macrophage progenitors (GMP)¹¹ and injected them into sublethally irradiated syngeneic recipient mice (Supplementary Fig. 1). Within 80 days, animals transplanted with murine stem cell virus (MSCV)-MLL–AF9 transduced GMP developed an oligoclonal acute myelogenous leukaemia (AML), similar to other MLL fusion leukaemia models (Supplementary Fig. 2 and Supplementary Table 1)^{9,12–15}. To assess the frequency of leukaemia-initiating cells, we injected sublethally irradiated syngeneic recipient mice with 20 to 10⁴ bone-marrow cells taken from leukaemic mice. All mice that received 500 or more leukaemic cells died of AML. Half of the mice that received 100 cells, and none of the mice that received 20 cells

succumbed to AML (Supplementary Table 1). Therefore, the calculated frequency of leukaemia-initiating cells in bone marrow is approximately 1:150. Thus MLL–AF9 induces self-renewal and leukaemia from prospectively isolated GMP.

Retroviral or knock-in models of MLL-fusion-induced leukaemias demonstrate the presence of IL-7R[–]Lin[–]Sca-1[–]c-Kit⁺CD34⁺FcγRII/III⁺ GMP-like leukaemic cells (L-GMP) (Fig. 1a)^{9,14}. We reasoned that these L-GMP might contain LSC. To determine whether the L-GMP population contains LSC, we sorted Lin⁺, IL-7R[–]Lin[–]Sca-1[–]c-Kit[–], or L-GMP cells from mice that developed AML (Fig. 1a). *Ex vivo* growth of these cells in cytokine-supplemented

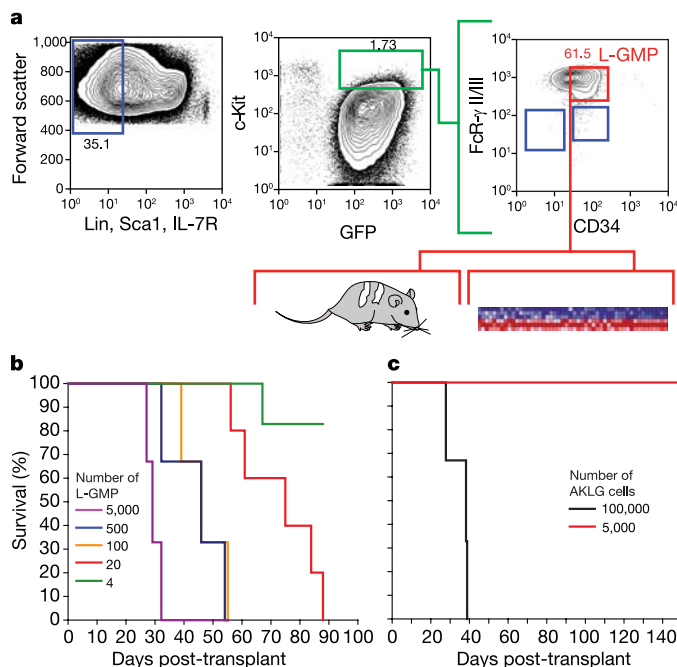


Figure 1 | Leukaemic-GMP are enriched for leukaemia stem cells.

a, Progenitor analysis of MLL–AF9 leukaemias. L-GMP were prospectively isolated for injection to secondary recipients or gene expression analysis. Further characterization of leukaemias can be found in Supplementary Figs 1 and 2. **b**, Survival curves for mice injected with limiting dilution of L-GMP show that four cells can transfer leukaemia. The experiment was repeated three times with similar results. Detailed numbers can be found in Supplementary Table 1. **c**, Survival curves for three mice injected with 5,000 or 100,000 L-GMP propagated in culture (AKLG cells) demonstrate the requirement for >5,000 cells to initiate leukaemia.

¹Division of Hematology/Oncology, Children's Hospital, ²Department of Pediatric Oncology, and ³Medical Oncology, Dana Farber Cancer Institute, and ⁴Division of Hematology, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA. ⁵The Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA. ⁶Howard Hughes Medical Institute, Chevy Chase, Maryland 20815, USA.

methylcellulose revealed that L-GMP possessed higher blast colony-forming activity than other populations (Supplementary Fig. 3). Transplantation of purified L-GMP into sublethally irradiated recipients was used to assess leukaemogenesis *in vivo*. Each of the mice transplanted with 20 to 5×10^3 L-GMP developed AML that was phenotypically identical to the primary disease (Fig. 1b, Supplementary Fig. 4). One of six mice injected with four L-GMP died of AML (Fig. 1b). Therefore, the L-GMP population contains leukaemia-initiating cells with a frequency of approximately 1:6.

In addition, a high frequency of leukaemia-initiating cells in the L-GMP was confirmed by plating in cytokine-supplemented methylcellulose. Single colonies were isolated after 5 days and injected into recipient mice. Fifteen out of 30 mice receiving cells from single L-GMP derived colonies died of AML within 70 days (data not shown). To further characterize the stem-cell-like properties of L-GMP, we assessed their potential to generate differentiated progeny. L-GMP were easily propagated in liquid culture containing IL-3, a well-described characteristic of murine MLL-fusion-induced leukaemias¹². However, after conversion to liquid culture the majority of the cells demonstrated a more differentiated, lineage⁺ Kit⁻ immunophenotype (Supplementary Fig. 5). This more differentiated phenotype was seen in conjunction with an increase in the number of cells needed to initiate leukaemia in secondary recipient mice to >5,000 (Fig. 1c). These data demonstrate that leukaemia cells with an immunophenotype similar to normal GMP are highly enriched for LSC that can both initiate leukaemia in secondary recipient mice and produce more differentiated progeny incapable of transferring the disease.

Having prospectively purified a population highly enriched for LSC, we used gene expression profiling to assess cellular identity, and characterize changes that occur during the transition from committed progenitor to LSC. Because isolation of L-GMP is dependent upon the expression of a limited number of immunophenotypic markers, we assessed whether the gene expression profile of L-GMP remained similar to the normal GMP from which they arose or whether global cellular reprogramming had occurred during the transformation. To compare gene expression profiles between normal progenitors and L-GMP, we isolated the IL7R⁻ Lin⁻ Sca-1⁺ c-kit⁺ haematopoietic stem cell (HSC)-enriched population, IL-7R⁻ Lin⁻ Sca-1⁻ c-Kit⁺ CD34⁺ FcγRII/III^{lo} common myeloid progenitors (CMP), IL-7R⁻ Lin⁻ Sca-1⁻ c-Kit⁺ CD34⁺ FcγRII/III^{hi} megakaryocyte erythroid progenitors (MEP), and GMP from 6–8-week-old C57BL/6 mice. We also isolated L-GMP from mice that were approximately 60 days post transplant and had clinically evident leukaemia. RNA was purified from each population, amplified and hybridized to Affymetrix murine 430A 2.0 microarrays. Unsupervised analysis demonstrated that the L-GMP have a stable gene expression programme that is different from any of the normal HSC or progenitor populations, but most similar to the GMP from which they arose (Fig. 2a). *K*-means clustering identified a large group of genes that demonstrate high-level expression in normal GMP as compared to other normal progenitors and HSCs, and expression that is further elevated in the L-GMP population (signature 1a in Fig. 2b). The inverse of this profile was similarly prominent (signature 1b in Fig. 2b).

An intriguing group of genes are highly expressed in the HSC population, show decreased expression in committed progenitors, but have re-activated high-level expression in the L-GMP (signature 3a in Fig. 2b), and the inverse (signature 3b in Fig. 2b). Supervised analysis identified 363 genes that were more highly expressed in populations with self-renewal potential (HSC and L-GMP) than in progenitors with no self-renewal potential ($P < 0.001$) (Fig. 2c), and 262 genes that are more highly expressed in committed progenitors than either HSC or L-GMP ($P < 0.001$) (Supplementary Fig. 6). We note that the 363 genes that are more highly expressed in HSC and L-GMP (a self-renewal-associated signature) represent only a subset of the approximately 1137 genes normally highly expressed in HSC

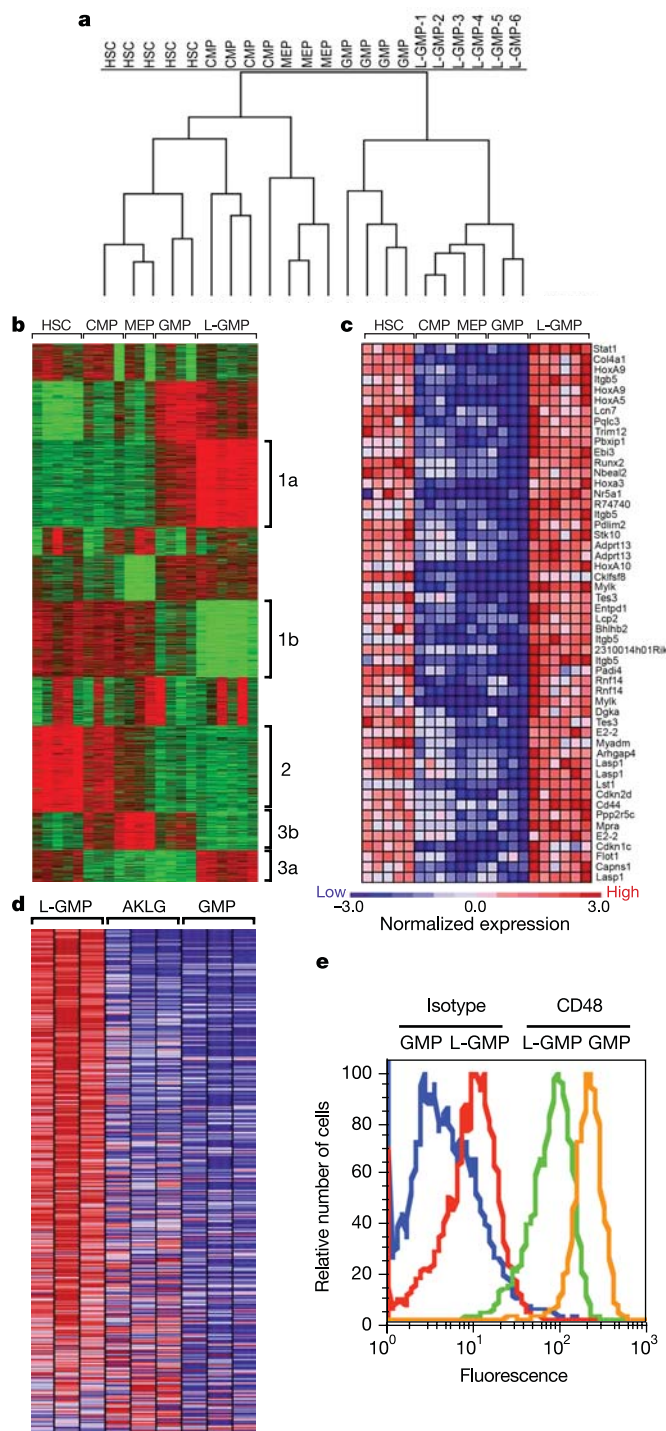


Figure 2 | Progenitor-derived leukaemia stem cells maintain progenitor identity and re-activate a self-renewal-associated programme.

a, Hierarchical clustering using the 9,100 filtered probe sets demonstrates that L-GMP possess a gene expression profile most similar to GMP. **b**, *K*-means clustering ($K = 10$ clusters specified) identifies the major gene expression clusters, labelled on the right. **c**, The top 50 probe sets for genes that show increased expression in the HSC population and L-GMP are shown. Permutation testing demonstrates 420 probe sets (363 genes) in this signature ($P < 0.001$). **d**, The 363-gene self-renewal-associated signature was assessed in L-GMP, AKLG cells, and GMP. GSEA demonstrates higher-level expression in the L-GMP ($P < 0.001$). **e**, CD48 expression on L-GMP and GMP as compared to an isotype control antibody.

compared to committed progenitors (Fig. 2b and Supplementary Fig. 7). As an independent validation of our HSC and self-renewal-associated signatures, we found that both were highly enriched ($P < 0.001$) in a previously published gene expression profile of long-term HSC¹⁶ (Supplementary Fig. 7). Next, we assessed whether the self-renewal-associated signature is lost upon propagation of the L-GMP in liquid culture under conditions that support robust *in vitro* growth, but leads to a decrease in the LSC frequency (Fig. 1c and Supplementary Fig. 5). When L-GMP differentiate into a largely non-self-renewing population (AKLG cells), the majority of the self-renewal-associated signature reverts back to that found in normal GMP, even though the cells remain immortal (Fig. 2d). Finally, the presence of a number of transcription factors that are important in normal haematopoietic development, and in leukaemogenesis mediated by MLL-fusions, including multiple *Hox* genes^{17–21}, supports the relevance of this signature (Fig. 2c). These data show that acquisition of self-renewal properties in committed progenitors is associated with activation of a portion of a programme highly expressed in normal HSC.

To assess further the homogeneity of this population, we searched for genes in the self-renewal-associated signature encoding cell-surface molecules that we might assess by flow cytometry. One

such gene, *CD48*, is a member of the Slam family of cell-surface molecules, shown to be part of a Slam-family signature in haematopoietic stem cells²². Our gene-expression data demonstrated that *CD48* is more highly expressed in GMP than in either HSC or L-GMP (signature 3b in Fig. 2b). Assessment of *CD48* expression on the surface of GMP and L-GMP demonstrated decreased and homogeneous expression in L-GMP (Fig. 2e). We further assessed homogeneity of the L-GMP using a computational approach described in Supplementary Fig. 8. These data show L-GMP is a homogeneous population, and demonstrate the potential utility of the self-renewal-associated signature for the identification of markers that might be used to further characterize LSC.

To assess the relevance of the murine signature in human AML, we analysed expression of the self-renewal-associated gene expression programme in human *MLL*-rearranged AML as compared to AML with other chromosomal translocations using a previously published human AML data set²³. Gene set enrichment analysis (GSEA)²⁴ demonstrated significant overlap between the murine self-renewal-associated signature and the human *MLL*-AML signature with 91/363 genes more highly expressed in *MLL*-rearranged AML ($P = 0.016$) (Fig. 3a). Thus, a portion of the self-renewal-associated programme is highly expressed in *MLL*-rearranged AML supporting

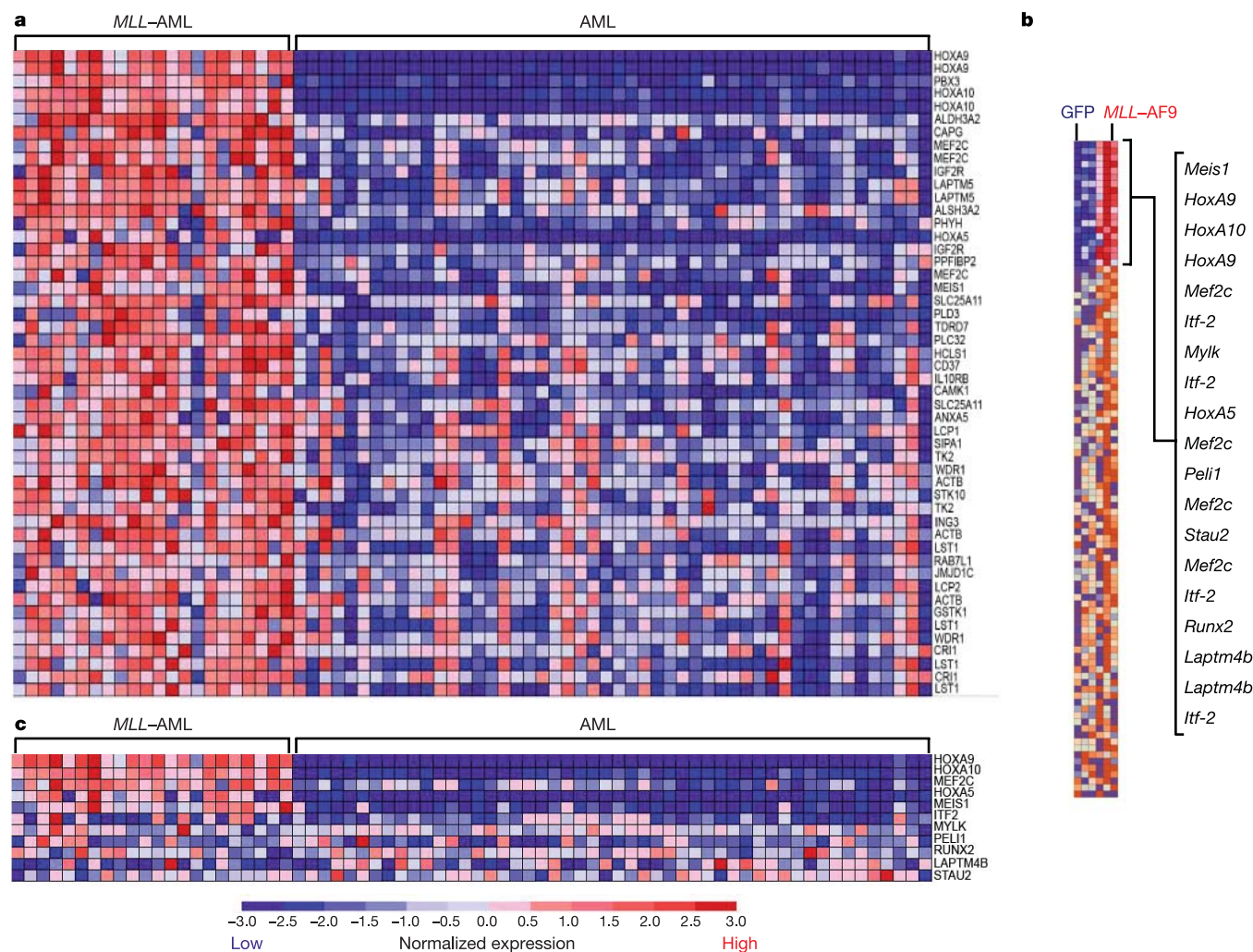


Figure 3 | The self-renewal-associated signature is found in human *MLL*-rearranged AML and activated as a hierarchy of gene expression. **a, GSEA demonstrated significant enrichment ($P = 0.016$) of the murine self-renewal-associated signature in human *MLL*-AML compared to other AML with defined translocations. **b**, The 420 probe sets from the self-renewal-associated signature were ranked according to the distinction of**

high-level expression in *MLL*-AF9-GFP transduced GMP compared to MSCV-GFP transduced GMP 40 h after incubation with retroviruses. Eleven genes had a t -test score > 2.0 and are labelled to the right. Genes with a score < 2.0 are shaded. **c**, The 11-gene set was assessed in *MLL*-rearranged human AMLs as in **a**. GSEA demonstrated significant overlap ($P = 0.004$).

the relevance of this signature in human AML. Future studies will determine whether the genes that are similarly expressed in all AML samples are part of a universal AML self-renewal signature.

Having identified a self-renewal-associated signature present in LSC, we determined whether it is fully activated immediately after MLL-AF9 expression, or whether there is a hierarchy of gene expression changes that might help us focus on individual genes. We transduced GMP with either MLL-AF9-green fluorescent protein (GFP) or MSCV-GFP retroviruses. Forty hours later GFP⁺ cells were re-sorted, RNA was isolated and hybridized to microarrays. We assessed changes in expression of genes in the self-renewal-associated signature and found that a small subset showed increased expression in the MLL-AF9-GFP transduced GMP (Fig. 3b). These 'immediate' genes included genes previously implicated in MLL-rearranged leukaemias *HoxA5*, *HoxA9*, *HoxA10*, *Meis1* and new genes not known to have an association with MLL-rearrangement such as *Mef2c*, *Runx2* and the *wnt*-pathway associated gene *Itf-2* (ref. 25). GSEA demonstrated significant enrichment ($P = 0.004$) of the immediate signature in human MLL-rearranged AML as compared to other AMLs with six of 11 genes in the signature more highly expressed in the MLL-AML (Fig. 3c). Thus, a subset of genes are immediately activated by MLL-AF9 expression suggesting a hierarchy of gene expression where the total self-renewal-associated signature is a summation of multiple gene expression programmes.

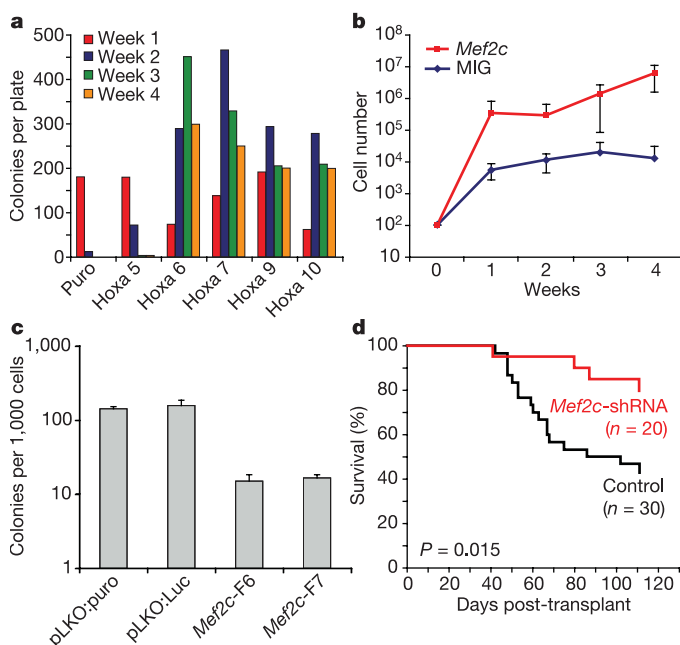


Figure 4 | *Hox* genes and *Mef2c* play a role in L-GMP. **a**, 5,000 GMP transduced with the indicated retroviruses were plated in cytokine-supplemented methylcellulose. Every seven days the number of colonies was determined, and 5,000 cells re-plated. **b**, GMP transduced with *Mef2c* were subjected to serial replating as above. The colony forming activity was unchanged, but there were consistently fivefold more cells in the *Mef2c*-transduced colonies as compared to control cells. Error bars are s.d. of duplicates. The experiment was repeated five times with similar results. **c**, L-GMP colonies grown in the presence of puromycin were counted on day 5 after transduction with either an empty lentivirus (pLKO:puro), or viruses expressing either luciferase, *Mef2c*-F6 or *Mef2c*-F7 shRNAs. Error bars are s.d. of duplicates. **d**, Colonies isolated day 5 after transduction with a control lentivirus or *Mef2c*-F6 shRNA were injected into recipient mice, and monitored for AML. A survival curve demonstrates a significant difference in survival when the experiment was terminated at day 112 ($P = 0.015$). On day 112, one mouse in each group contained GFP-positive cells and were thus counted as having developed AML.

Next we focused attention on genes both immediately and highly expressed in murine and human MLL-AML (Fig. 3c). Given the prominent role *HoxA* cluster genes play in MLL-rearranged and other leukaemias²⁶, and the fact that MLL may regulate chromatin structure near the 5' *HoxA* cluster^{27,28}, we assessed which of the 5' *HoxA* cluster genes could impart serial replating activity on GMP. We expressed *HoxA5*-*A10* in GMP, and found that *HoxA6*, *HoxA7*, *HoxA9* or *HoxA10*, but not *HoxA5*, were capable of inducing serial replating activity (Fig. 4a). Also, the tight correlation between MLL-AF9 and *Mef2c* expression prompted a similar assessment. Expression of *Mef2c*, a member of the myocyte enhancer factor 2 gene family of transcription factors, in GMP consistently produced a 5–10-fold increase in the number of differentiated myelomonocytic cells, but did not induce serial replating activity (Fig. 4b). This observation is consistent with a recent study that identified *Mef2c* as a cooperating oncogene in leukaemogenesis that is unable to induce leukaemia when expressed alone²⁹. Next, we expressed two *Mef2c*-directed, or control short hairpin (sh)RNAs in L-GMP. The *Mef2c*-shRNAs suppressed *Mef2c* expression (Supplementary Fig. 9), and decreased colony number and colony size (Fig. 4c). Finally, we isolated individual colonies grown from L-GMP transduced with a control virus or a virus expressing the *Mef2c*-F6 shRNA, and injected them into sublethally irradiated syngeneic recipient mice. Seventeen of 30 (56%) mice injected with control L-GMP colonies developed a lethal disease consistent with AML. However, only 20% (four of 20) of mice injected with *Mef2c*-F6 transduced L-GMP developed AML (Fig. 4d). These data support a role for *Hox* genes and *Mef2c* in leukaemia stem cells, and provide evidence that genes in the self-renewal-associated signature are important for LSC development.

We have shown that progenitor-derived LSC are globally most similar to the progenitor from which they arose. However, the LSC express a self-renewal-associated programme normally expressed in HSC (Supplementary Fig. 10). These data suggest that ectopic reactivation of genes associated with self-renewal is a central feature of leukaemic transformation of progenitor cells. Further characterization should determine which genes are involved in LSC self-renewal or immortalization and the extent to which these processes overlap. Future studies will assess whether LSC, but not HSC, are dependent upon specific components of this programme, and thus might be successfully targeted therapeutically. If LSC-specific targeting can be achieved, this should complement existing therapies by providing activity against the subpopulation of tumour cells with the unique capacity for self-renewal.

METHODS

See the Supplementary Methods for detailed experimental methods.

Mice, antibodies, and sorting of haematopoietic progenitors. 8–10-week-old C57BL/6 mice (Charles River Laboratories) were used as bone-marrow donors and recipients. Myeloid progenitors were sorted as previously described¹¹.

Retroviruses and culture of haematopoietic progenitors. Retroviral supernatants were produced by co-transfection of 293T cells³⁰. shRNA in lentiviral vectors were obtained from the RNAi Consortium, and viral particles were generated by co-transfection of 293T cells with lentiviral packaging plasmids. 5×10^4 to 5×10^5 GMP were incubated with retroviral supernatant, and 40 h after infection re-sorted GFP⁺ PI-negative cells were either injected into sublethally irradiated recipient mice or collected for RNA. L-GMP sorted from leukaemic mice were injected into syngeneic recipient mice or incubated with lentiviral vectors and plated in methylcellulose media M3234 or liquid culture (Stem Cell Technologies) supplemented with 10 ng ml⁻¹ IL-3 (Peprotech), 1× penicillin/streptomycin (Gibco), and 2.5 µg ml⁻¹ puromycin (Sigma).

RNA amplification and qPCR. We routinely isolated RNA from $(2-5) \times 10^4$ cells, and performed two rounds of *in vitro* transcription and labelling as described¹⁴ (<http://www.broad.mit.edu/mpr/publications/projects/leukemia/protocol.html>). The quantitative polymerase chain reaction (qPCR) is described in the Supplementary Methods.

Data analysis and statistical methods. Data were analysed with either GSEA, Gene Pattern (<http://www.broad.mit.edu/tools/software.html>), or Cluster software (<http://rana.lbl.gov/EisenSoftware.htm>).

Received 9 March; accepted 14 June 2006.

Published online 16 July 2006.

1. Reya, T., Morrison, S. J., Clarke, M. F. & Weissman, I. L. Stem cells, cancer, and cancer stem cells. *Nature* **414**, 105–111 (2001).
2. Warner, J. K., Wang, J. C., Hope, K. J., Jin, L. & Dick, J. E. Concepts of human leukemic development. *Oncogene* **23**, 7164–7177 (2004).
3. Al-Hajj, M. & Clarke, M. F. Self-renewal and solid tumor stem cells. *Oncogene* **23**, 7274–7282 (2004).
4. Bonnet, D. & Dick, J. E. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nature Med.* **3**, 730–737 (1997).
5. Lapidot, T. *et al.* A cell initiating human acute myeloid leukaemia after transplantation into SCID mice. *Nature* **367**, 645–648 (1994).
6. Singh, S. K. *et al.* Identification of human brain tumour initiating cells. *Nature* **432**, 396–401 (2004).
7. Jamieson, C. H. *et al.* Granulocyte-macrophage progenitors as candidate leukemic stem cells in blast-crisis CML. *N. Engl. J. Med.* **351**, 657–667 (2004).
8. Huntly, B. J. *et al.* MOZ-TIF2, but not BCR-ABL, confers properties of leukemic stem cells to committed murine hematopoietic progenitors. *Cancer Cell* **6**, 587–596 (2004).
9. Cozzio, A. *et al.* Similar MLL-associated leukemias arising from self-renewing stem cells and short-lived myeloid progenitors. *Genes Dev.* **17**, 3029–3035 (2003).
10. So, C. W. *et al.* MLL-GAS7 transforms multipotent hematopoietic progenitors and induces mixed lineage leukemias in mice. *Cancer Cell* **3**, 161–171 (2003).
11. Akashi, K., Traver, D., Miyamoto, T. & Weissman, I. L. A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature* **404**, 193–197 (2000).
12. Lavau, C., Szilvassy, S. J., Slany, R. & Cleary, M. L. Immortalization and leukemic transformation of a myelomonocytic precursor by retrovirally transduced HRX-ENL. *EMBO J.* **16**, 4226–4237 (1997).
13. Corral, J. *et al.* An MLL-AF9 fusion gene made by homologous recombination causes acute leukemia in chimeric mice: a method to create fusion oncogenes. *Cell* **85**, 853–861 (1996).
14. Wang, J. *et al.* Conditional MLL-CBP targets GMP and models therapy-related myeloproliferative disease. *EMBO J.* **24**, 368–381 (2005).
15. Drynan, L. F. *et al.* Mll fusions generated by Cre-loxP-mediated de novo translocations can induce lineage reassignment in tumorigenesis. *EMBO J.* **24**, 3136–3146 (2005).
16. Ramalho-Santos, M., Yoon, S., Matsuzaki, Y., Mulligan, R. C. & Melton, D. A. "Stemness": transcriptional profiling of embryonic and adult stem cells. *Science* **298**, 597–600 (2002).
17. Zeisig, B. B. *et al.* Hoxa9 and Meis1 are key targets for MLL-ENL-mediated cellular immortalization. *Mol. Cell. Biol.* **24**, 617–628 (2004).
18. Kumar, A. R. *et al.* Hoxa9 influences the phenotype but not the incidence of Mll-AF9 fusion gene leukemia. *Blood* **103**, 1823–1828 (2004).
19. Ayton, P. M. & Cleary, M. L. Transformation of myeloid progenitors by MLL oncoproteins is dependent on Hoxa7 and Hoxa9. *Genes Dev.* **17**, 2298–2307 (2003).
20. Armstrong, S. A. *et al.* MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genet.* **30**, 41–47 (2002).
21. So, C. W., Karsunky, H., Wong, P., Weissman, I. L. & Cleary, M. L. Leukemic transformation of hematopoietic progenitors by MLL-GAS7 in the absence of Hoxa7 or Hoxa9. *Blood* **103**, 3192–3199 (2004).
22. Kiel, M. J., Yilmaz, O. H., Iwashita, T., Terhorst, C. & Morrison, S. J. SLAM family receptors distinguish hematopoietic stem and progenitor cells and reveal endothelial niches for stem cells. *Cell* **121**, 1109–1121 (2005).
23. Ross, M. E. *et al.* Gene expression profiling of pediatric acute myelogenous leukemia. *Blood* **104**, 3679–3687 (2004).
24. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
25. Kolligs, F. T. *et al.* ITF-2, a downstream target of the Wnt/TCF pathway, is activated in human cancers with beta-catenin defects and promotes neoplastic transformation. *Cancer Cell* **1**, 145–155 (2002).
26. Abramovich, C. & Humphries, R. K. Hox regulation of normal and leukemic hematopoietic stem cells. *Curr. Opin. Hematol.* **12**, 210–216 (2005).
27. Milne, T. A. *et al.* MLL associates specifically with a subset of transcriptionally active target genes. *Proc. Natl Acad. Sci. USA* **102**, 14765–14770 (2005).
28. Guenther, M. G. *et al.* Global and Hox-specific roles for the MLL1 methyltransferase. *Proc. Natl Acad. Sci. USA* **102**, 8603–8608 (2005).
29. Du, Y., Spence, S. E., Jenkins, N. A. & Copeland, N. G. Cooperating cancer-gene identification through oncogenic-retrovirus-induced insertional mutagenesis. *Blood* **106**, 2498–2505 (2005).
30. Ernst, P., Mabon, M., Davidson, A. J., Zon, L. I. & Korsmeyer, S. J. An Mll-dependent Hox programme drives hematopoietic progenitor expansion. *Curr. Biol.* **14**, 2063–2069 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank B. Huntly and A. Kung for discussions and ViaCell Inc. for the use of a high-speed cell sorter in the early stages of the project. We thank members of the Akashi laboratory for instruction regarding initial progenitor sorting experiments. We also thank E. Schindler and E. Smith for assistance. This work was funded by the National Institutes of Health, the Leukemia Lymphoma Society, the Richard and Susan Smith foundation, and the Damon Runyon Cancer Research Foundation (to S.A.A.). S.A.A. is a Damon Runyon Clinical Investigator. We dedicate this work to the late Stanley J. Korsmeyer, whose scientific vision inspired the planning of this project.

Author Information The microarray data discussed here can be found in the Gene Expression Omnibus (GEO) of NCBI at <http://www.ncbi.nlm.nih.gov/geo/> through accession numbers GSE3725 and GSE4416 or at http://www.broad.mit.edu/cgi-bin/cancer/publications/pub_menu.cgi. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.A.A. (Scott.Armstrong@childrens.harvard.edu).

Notch signalling regulates stem cell numbers *in vitro* and *in vivo*

Andreas Androutsellis-Theotokis¹, Ronen R. Leker¹, Frank Soldner¹, Daniel J. Hoepfner¹, Rea Ravin¹, Steve W. Poser¹, Maria A. Rueger¹, Soo-Kyung Bae¹, Raja Kittappa¹ & Ronald D. G. McKay¹

The hope of developing new transplantation therapies for degenerative diseases is limited by inefficient stem cell growth and immunological incompatibility with the host^{1,2}. Here we show that Notch receptor activation induces the expression of the specific target genes *hairy* and *enhancer of split 3 (Hes3)* and *Sonic hedgehog (Shh)* through rapid activation of cytoplasmic signals, including the serine/threonine kinase Akt, the transcription factor STAT3 and mammalian target of rapamycin, and thereby promotes the survival of neural stem cells. In both murine somatic and human embryonic stem cells, these positive signals are opposed by a control mechanism that involves the p38 mitogen-activated protein kinase. Transient administration of Notch ligands to the brain of adult rats increases the numbers of newly generated precursor cells and improves motor skills after ischaemic injury. These data indicate that stem cell expansion *in vitro* and *in vivo*, two central goals of regenerative medicine, may be achieved by Notch ligands through a pathway that is fundamental to development and cancer^{3–5}.

New cell therapies based on embryonic stem (ES) cells are supported by work in animal models of human disease. They are difficult to implement, however, because it is hard to grow tissue-specific precursors in the laboratory and it is difficult to deliver them to diffuse disease sites in the body without stimulating an immune response. The results that we present here suggest a general model of stem cell expansion that applies to many precursor cells of clinical interest and that may lead to strategies that promote regenerative responses through activation of endogenous cells (Supplementary Fig. 1).

Notch encodes a transmembrane receptor that is cleaved to release an intracellular domain (Nictd) that is directly involved in transcriptional control^{3,6}. Many components of the Notch pathway are expressed in the precursor cell compartment of the developing vertebrate central nervous system (CNS)^{7,8}. Here, cell culture conditions that support the growth of a homogeneous population of fetal neural stem cells (NSCs) *in vitro* were used to define the action of Notch ligands, Delta-like 4 (Dll4) and Jagged 1 (Jag1)⁹. Reaction with an antibody specific for the cleaved cytoplasmic fragment showed proteolysis of the transmembrane form of the Notch receptor (Fig. 1a and Supplementary Fig. 2a). When cells were continuously observed in a sealed chamber on a microscope stage, Dll4 rapidly reduced cell death (Fig. 1b and Supplementary Fig. 2b). This survival effect of Notch ligands was antagonized by DAPT, a γ -secretase inhibitor that blocks Notch cleavage by Presenilin 1 (Fig. 2a). Cells exposed to Notch ligands retained the potential to generate neurons, astrocytes and oligodendrocytes after prolonged exposure to Notch ligands (Supplementary Table 1). These data show that activation of the Notch receptor rapidly inhibits cell death in NSCs.

The rapid effect of Dll4 on NSC survival suggested that cytoplasmic

survival signals were induced in addition to slower transcriptional responses traditionally attributed to Notch activation. The insulin receptor stimulates cell survival through a cytoplasmic phosphorylation cascade that is initiated by phosphatidylinositol-3-OH kinase (PI(3)K) and the serine/threonine kinase Akt^{4,10,11}. Dll4 induced DAPT-dependent phosphorylation of two principal activation sites on Akt with a peak at 5 min (Fig. 1c, d). Downstream of Akt, mammalian target of rapamycin (mTOR) is a key regulator of cell growth¹². Jag1 caused transient phosphorylation of mTOR (Supplementary Fig. 2c). Like DAPT, the mTOR inhibitor rapamycin blocked the survival effect of Dll4 (Fig. 2a). These results suggest that Notch cleavage activates the survival cascade downstream of the insulin receptor.

STAT3 is a transcription factor activated at the cell surface by the gp130/LIF receptor and the JAK tyrosine kinase⁵. Extracellular ligands that promote phosphorylation of Tyr 705 on STAT3 stimulate the differentiation of CNS stem cells to a glia fate¹³. In our work on the role of STAT3 in fate choice, we noted that phosphorylation of a serine at position 727 correlated with the survival of NSCs. Dll4 and Jag1 induced dose-dependent phosphorylation of STAT3 on Ser 727 in the absence of Tyr 705 phosphorylation. Phosphorylation of Ser 727 peaked at 20 min, consistent with the kinetics of mTOR activation (Fig. 1e and Supplementary Fig. 2d). Ser 727 phosphorylation after Notch activation was sensitive to DAPT and rapamycin (Supplementary Fig. 2e). DAPT also reduced the basal amounts of phosphorylated Ser 727, suggesting that endogenous γ -secretase has a role in Ser 727 phosphorylation. These results suggest that serine phosphorylation of STAT3 mediates the survival effects of Notch activation.

We used a genetic approach to assess further the role of Ser 727 in cell survival pathways. NSCs were transfected with wild-type STAT3 and two mutant forms in which Tyr 705 or Ser 727 was altered to a 'neutral' amino acid (STAT3-YF and STAT3-SA). A gene encoding green fluorescent protein (GFP) in the plasmids enabled the number of transfected cells to be measured at 24 h and 4 d. The initial transfection efficiency was similar with all three complementary DNAs (~50% GFP⁺ cells at 24 h). At 4 d, there were similar numbers of wild-type and STAT3-YF transfected cells (Fig. 1f; wild-type STAT3, 100 ± 29%; STAT3-YF, 100 ± 27%) but the proportion of STAT3-SA transfected cells was greatly reduced (16 ± 11%). This result suggests that endogenous ligands stimulate phosphorylation of STAT3 on Ser 727. Similar results were obtained in the presence of Dll4 (wild-type STAT3, 100 ± 29%; STAT3-YF, 109 ± 23%; STAT3-SA, 20 ± 3%). These data show that Notch ligands achieve their survival effects only when Ser 727 of STAT3 can be phosphorylated.

Phosphorylation on Tyr 705 is thought to control transcription of the gene encoding STAT3 and also mediates cell survival in cancer cells^{5,14}. A form of STAT3 modified to prevent tyrosine

¹Laboratory of Molecular Biology, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, Maryland 20892, USA.

phosphorylation and to have constitutively active serine phosphorylation was used to confirm a specific survival role for this target of the Notch receptor pathway. Transfection with a plasmid encoding a Ser 727 phosphomimetic and Tyr705 → Phe double mutant (STAT3-YF/SE) significantly increased colony formation (Fig. 1f; STAT3-YF 100 ± 20%, STAT3-YF or STAT3-SE 305 ± 58%), suggesting that Ser 727 has a positive role in NSC survival.

The short duration of mTOR activation suggests the presence of a negative regulatory mechanism that limits the extent of the positive signal. The PDK1 and p70 ribosomal S6 kinase components of the insulin signalling pathway are known to limit mTOR activation through the MSK1 and LKB1 kinases, which have been intensively studied as drug targets in diabetes and cancer^{15,16}. Jag1 induced phosphorylation of MSK1 and LKB1 at 1 h after treatment, consistent with the presence of an inhibitory mechanism (Supplementary Fig. 2f).

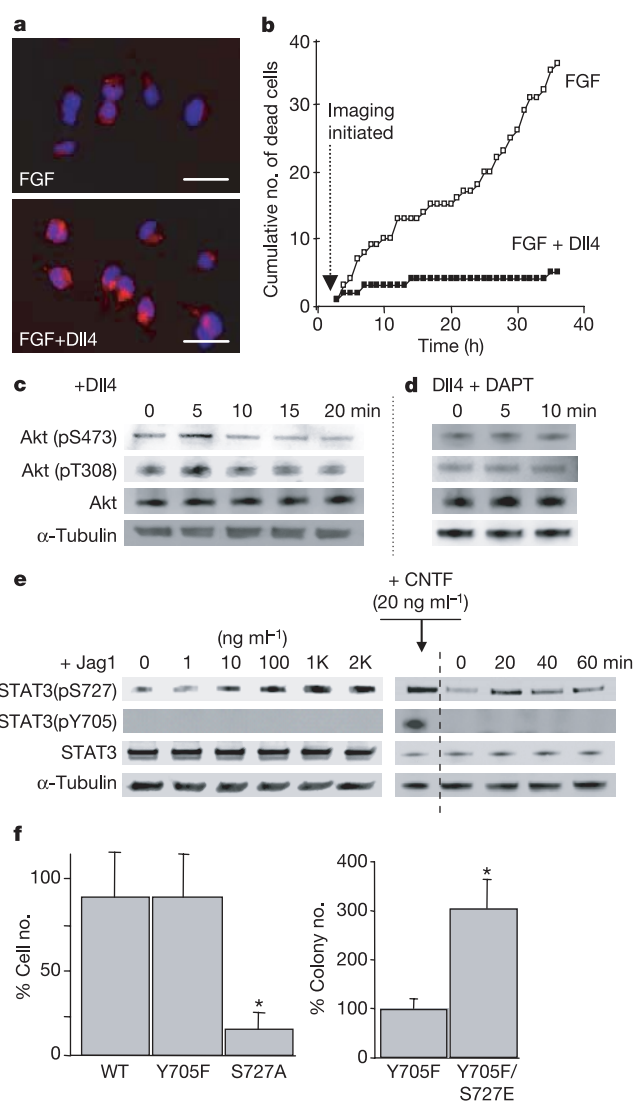


Figure 1 | Notch ligands activate second messenger signalling and support NSC (E13.5) survival *in vitro*. **a**, Dll4 (500 ng ml⁻¹, 1-h treatment) activates Notch1 (red, cleaved Notch1; blue, DAPI; scale bars 50 μm). **b**, Dll4 inhibits cell death in real-time cell lineage experiments. **c**, **d**, Dll4 activates Akt, with a peak at 5 min, in a DAPT-dependent manner. **e**, Jag1 induces phosphorylation of STAT3 on Ser 727 in a dose- and time-dependent manner (CNTF, positive control). **f**, Transfection with STAT3-YF does not alter survival relative to wild-type STAT3, whereas STAT3-SA inhibits growth and STAT3-YF/SE promotes survival (mean ± s.d. data 4 d after transfection).

The p38 mitogen-activated protein kinase is also a potential inhibitor of survival because it acts downstream of JAK and antagonizes growth in many cell types by activating MSK^{17,18}. JAK and p38 inhibitors increased survival in NSCs (Fig. 2a). Combined JAK and p38 inhibition did not substantially improve survival, further indicating that JAK may act through p38 to antagonize the survival pathway in NSC. These data suggest that Notch acting through STAT3 promotes, and that p38 antagonizes, survival.

These results raise the issue of whether the cytoplasmic functions of the Notch receptor regulate transcription of the *Hes/Hey* gene family. Jag1 induced rapid transcriptional activation of *Hes3*, a gene with roles in NSC maintenance *in vivo*⁸. The regulation of *Hes3* messenger RNA by Jag1 was blocked when JAK kinase was activated by ciliary neurotrophic factor (CNTF) (20 ng ml⁻¹); this block was overcome by a JAK inhibitor. Induction of *Hes3* mRNA was also blocked by DAPT and rapamycin (Fig. 2b). *Hes3* mRNA and Hes3 protein are both enriched in cells of the adult subventricular zone (SVZ), a restricted site in the adult CNS that sustains stem cells¹⁹ (Fig. 2c). Sonic hedgehog (Shh) is a secreted protein with mitogenic functions in precursor cells in the developing brain and a is chief target for treatments of paediatric brain tumours^{20,21}. Jag1 induced long-term expression of Shh protein (Fig. 2d); cDNA transfection demonstrated that this effect is downstream of Hes3 (Fig. 2e). Previous studies show that Notch promotes NSC survival and that Nid1 interacts with JAK and STAT^{22,23}. The data presented here suggest a model in which Notch ligands act through a series of rapid and precisely timed cytoplasmic responses that have been widely studied in the context of cancer (Fig. 3).

It is important to establish whether this control mechanism applies

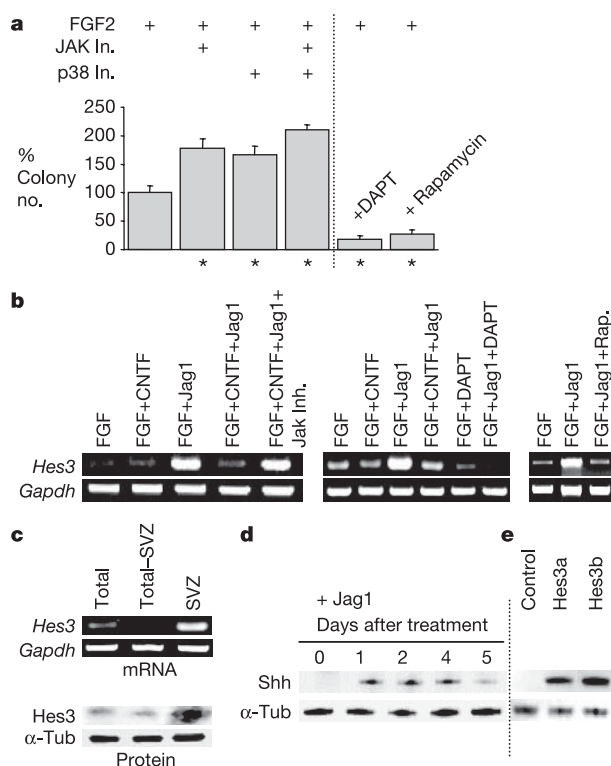


Figure 2 | Mediators and modulators of Notch signalling in NSC (E13.5) cultures. **a**, Inhibitors of JAK and p38 promote CNS stem cell survival; inhibitors of γ-secretase and mTOR inhibit survival (mean ± s.d. data 5 d after plating). **b**, CNTF (20 ng ml⁻¹) inhibits the Jag1-induced upregulation of *Hes3* mRNA; this effect is sensitive to a JAK inhibitor (200 nM). *Hes3* regulation by Jag1 is also rapamycin sensitive. **c**, *Hes3* mRNA and Hes3 protein are concentrated in the SVZ of adult mouse brains. **d**, Jag1 induces Shh expression in a time-dependent manner. **e**, Transfection with Hes3a or Hes3b induces Shh expression.

generally to ES and somatic stem cells. The clinical value of human ES (hES) cells is widely discussed but these cells are difficult to maintain in culture. Medium conditioned by mouse embryonic fibroblasts (MEFs) promotes their expansion. Although Akt and STAT3 have important roles in mouse ES cells, a clear function for STAT3 has not been defined in hES cells^{24,25}. Phosphorylation of STAT3 on Tyr 705 was undetectable in hES cells, but MEF conditioned medium maintained Ser 727 phosphorylation that was insensitive to JAK inhibition (Fig. 4a). High concentrations of fibroblast growth factor 2 (FGF2) and bone morphogenetic protein inhibitors are currently used to support human ES cell growth^{2,26}. Notch activation and inhibition of JAK and p38 (daily for 7 d) increased colony formation without affecting colony size in HSF6 cells (Fig. 4b). DAPT reduced colony number (control, $100 \pm 5\%$; DAPT $51 \pm 11\%$). Inhibition of JAK also increased survival in H1 and H9 hES cells (Supplementary Table 2).

To assess whether JAK inhibition was consistent with longer-term maintenance of the undifferentiated state, HSF6 cells were continuously exposed to the JAK inhibitor by daily additions for three passages (3 weeks). The cells had a normal morphology and antigen profile (Oct3/4⁺/SSEA4⁺/Tra-1-60⁺/Tra-1-81⁺/SSEA1⁻), and retained their ability to generate Sox1⁺/Nestin⁺ neural precursors and tyrosine-hydroxylase-positive neurons (data not shown). Similar results were obtained with endocrine precursors when cells from the developing pancreas were placed in culture (Fig. 4c, d). These data suggest that common regulators control the *in vitro* expansion of pluripotent and somatic stem cells. These cytoplasmic responses to Notch receptor activation may also contribute to the developmental inhibitory functions of Notch defined by genetic experiments in *Drosophila*⁶.

To determine whether Notch ligands are active *in vivo*, osmotic pumps were used to administer artificial cerebrospinal fluid (ACSF), FGF2, Dll4, or a combination of FGF2 plus Dll4 (Dll4, $4.2 \mu\text{g ml}^{-1}$; FGF2, $2.5 \mu\text{g ml}^{-1}$; $n = 6$ per treatment) for 7 d to the normal adult brain. Twice-daily injections of 5-bromodeoxyuridine (BrdU) from days 2 to 6 were given to label dividing cells. Dll4 increased the number of BrdU-positive cells in both the ipsi- and contralateral SVZ in a dose-dependent manner (Supplementary Fig. 3a). By contrast, FGF2 was ineffective (Supplementary Fig. 3b). Many newly generated cells expressed doublecortin, a protein that identifies immature neurons²⁷ (Supplementary Fig. 3c). Cells generated during the days after Notch ligand delivery were found at 45 d in the cerebral cortex; these cells were largely GFAP-negative and rarely (<1%) expressed the neuronal markers NeuN and calretinin (data not shown); however, many BrdU-positive cells in the cerebral cortex expressed the early neuronal marker HU (ipsilateral, $13.95 \pm 0.46\%$; contralateral, $10.2 \pm 2.35\%$; Supplementary Fig. 3d). These data show that

cells stimulated to divide by Dll4 survive for long periods in the parenchyma of the normal brain in an immature state. This result was seen in the absence of injury, indicating that Notch signalling may regulate the activity or size of the stem cell compartment in uncompromised adult tissues.

An injury model in which focal ischaemia was restricted to the cerebral cortex was used to determine whether Notch ligands promote regenerative responses. The middle cerebral artery was occluded at the surface of the brain to generate an injury restricted to the cerebral cortex. Exogenous proteins were delivered by osmotic pump, and standardized tests were used to measure motor behaviour over a 45-d survival period²⁸. In this injury model, Dll4 and FGF2 increased the numbers of BrdU-positive cells in both the ipsi- and

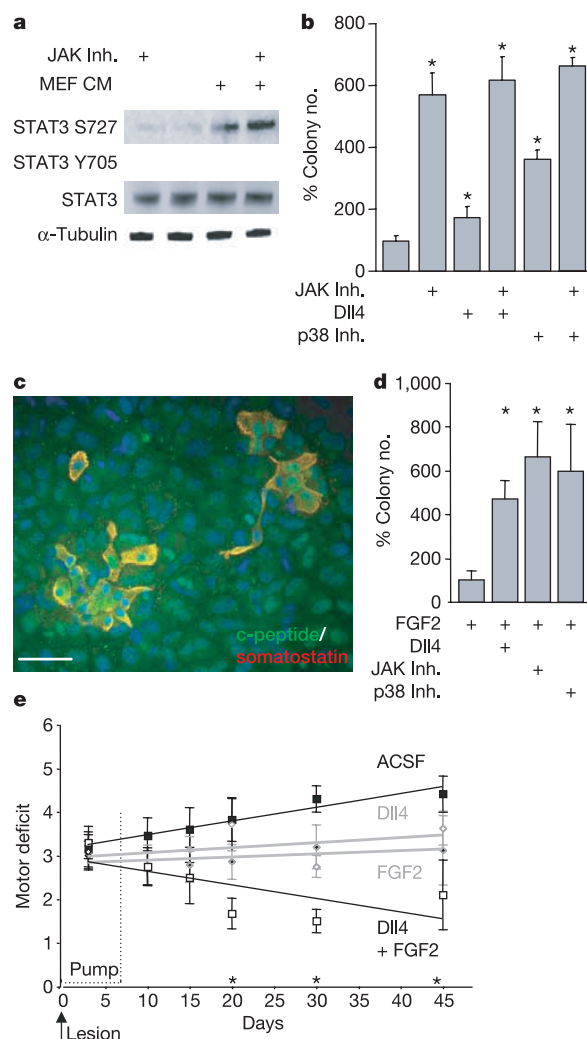


Figure 4 | *In vitro* and *in vivo* data support a general role for the signalling model in stem cell biology. **a**, MEF conditioned medium (MEF CM) stimulates phosphorylation of STAT3 on Ser 727: withdrawal of MEF CM (16 h) causes almost complete loss of Ser 727 phosphorylation without affecting STAT3 protein amounts in MEF-free HES cultures plated as aggregates; phosphorylation of STAT3 on Tyr 705 is undetectable in all conditions shown. **b**, Dll4, JAK inhibitor and p38 inhibitor increase plating efficiency of HSF6 hES cells plated as single dissociated cells (mean $\pm$ s.d. values 6 d after plating). **c**, Dll4 supports pancreatic endocrine precursor cultures expressing c-peptide and somatostatin (7 d after plating). **d**, Dll4, JAK inhibitor and p38 inhibitor promote survival of fetal pancreatic cultures (data are mean $\pm$ s.d. at 7 d). **e**, Treatment with Dll4 ($42 \mu\text{g ml}^{-1}$ in pump) plus FGF2 ($2.5 \mu\text{g ml}^{-1}$ in pump) improves motor skills of ischaemic rats (regression lines assume linear rate of change over the whole period of observation; data are mean $\pm$ s.e.m.).

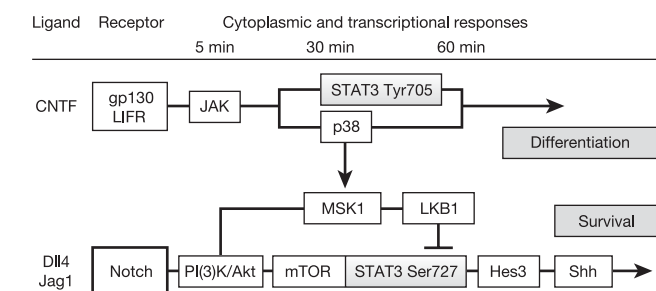


Figure 3 | A stem cell signalling pathway that affects survival and differentiation. Notch activation is followed by phosphorylation of Akt, mTOR and STAT3 on Ser 727, and by subsequent induction of Hes3 and Shh, in a temporally controlled order. Notch cleavage also simultaneously activates an opposing pathway that limits growth through MSK1 and LKB1. This opposing pathway can be accessed by gp130-dependent ligands that activate JAK and p38.

contralateral SVZ (Supplementary Fig. 3e). Here again, very few of the BrdU-labelled cells expressed markers characteristic of differentiated neurons or astrocytes. In lesioned rats treated with ACSF, there was a progressive deterioration in motor performance (Fig. 4e). The groups treated with either FGF or Dll4 alone showed no change in their motor scores over 45 d. Rats treated with both FGF2 and Dll4 showed significant improvement in motor skills. The size of the lesion was comparable in all treatment groups ($21.3 \pm 2.5\%$, ACSF; $21.0 \pm 3.5\%$, FGF2; $20.5 \pm 2.9\%$, FGF2 + Dll4; % of hemisphere volume at 7 d). The lack of an immediate neuroprotective effect suggests the exogenous proteins do not have a strong short-term anti-inflammatory function²⁹. Notch is important in all tissues and the beneficial effects of Notch ligands may involve responses in cells of the vascular, immune and nervous systems. *In vitro* data suggest that the Notch receptor has an important role in responses to low-oxygen conditions³⁰. The powerful *in vivo* effect that we have shown here indicates that Notch ligands may stimulate regenerative responses to the oxygen deprivation that follows ischaemia *in vivo*. Further studies will define how interactions between Notch ligands and other growth factors can be optimized to obtain more complete differentiation and recovery.

METHODS

Full methods are given in the Supplementary Information.

Cell culture. CNS stem cells and pancreatic precursors were cultured from mouse embryos at embryo day 13.5 (E13.5). Human ES cell lines (HSF6, H1, H9) were maintained on MEFs according to the suppliers' protocols.

Reverse transcriptase PCR. Newly designed and previously published (a gift from R. Kageyama, Institute for Virus Research, Kyoto University, Japan) primers were used (see Supplementary Information).

Nucleofection. STAT3 (a gift from T. Kitamura, Institute of Medical Sciences, Tokyo, Japan) and Hes3 (a gift from R. Kageyama) plasmid DNAs were expressed in CNS stem cells using the VPG-1004 nucleofector kit (Amaxa).

In vivo experiments. Drugs were administered over 7 d by an osmotic pump (Azlet). Rats received intraperitoneal injections of BrdU (50 mg per kg) every 12 h for 5 d beginning on day 1 after the operation. For the induction of focal ischaemia in rats, male spontaneously hypertensive rats underwent permanent middle cerebral artery occlusion²⁸.

Statistical analysis. Results shown are the mean $\pm$ s.d. or mean $\pm$ s.e.m. as indicated. Asterisks identify experimental groups that were significantly different from control groups by the Student's *t*-test (Microsoft Excel), with a Bonferroni correction for multiple comparisons (α -value, 0.05), where applicable.

Reagents. Reagents are listed in Supplementary Information.

Received 14 February; accepted 19 May 2006.

Published online 25 June 2006.

- Wakayama, T. *et al.* Differentiation of embryonic stem cell lines generated from adult somatic cells by nuclear transfer. *Science* **292**, 740–743 (2001).
- Xu, R. H. *et al.* Basic FGF and suppression of BMP signaling sustain undifferentiated proliferation of human ES cells. *Nature Methods* **2**, 185–190 (2005).
- Artavanis-Tsakonas, S., Rand, M. D. & Lake, R. J. Notch signaling: cell fate control and signal integration in development. *Science* **284**, 770–776 (1999).
- Cantley, L. C. The phosphoinositide 3-kinase pathway. *Science* **296**, 1655–1657 (2002).
- Levy, D. E. & Darnell, J. E. Jr Stats: transcriptional control and biological impact. *Nature Rev. Mol. Cell Biol.* **3**, 651–662 (2002).
- Goriely, A., Dumont, N., Dambly-Chaudière, C. & Ghysen, A. The determination of sense organs in *Drosophila*: effect of the neurogenic mutations in the embryo. *Development* **113**, 1395–1404 (1991).
- Tokunaga, A. *et al.* Mapping spatio-temporal activation of Notch signaling during neurogenesis and gliogenesis in the developing mouse brain. *J. Neurochem.* **90**, 142–154 (2004).
- Hatakeyama, J. *et al.* Hes genes regulate size, shape and histogenesis of the nervous system by control of the timing of neural stem cell differentiation. *Development* **131**, 5539–5550 (2004).
- Johe, K. K., Hazel, T. G., Muller, T., Dugich-Djordjevic, M. M. & McKay, R. D. Single factors direct the differentiation of stem cells from the fetal and adult central nervous system. *Genes Dev.* **10**, 3129–3140 (1996).
- Franke, T. F. *et al.* The protein kinase encoded by the Akt proto-oncogene is a target of the PDGF-activated phosphatidylinositol 3-kinase. *Cell* **81**, 727–736 (1995).
- Dudek, H. *et al.* Regulation of neuronal survival by the serine-threonine protein kinase Akt. *Science* **275**, 661–665 (1997).
- Nave, B. T., Ouwens, M., Withers, D. J., Alessi, D. R. & Shepherd, P. R. Mammalian target of rapamycin is a direct target for protein kinase B: identification of a convergence point for opposing effects of insulin and amino-acid deficiency on protein translation. *Biochem J.* **344**, 427–431 (1999).
- Rajan, P. & McKay, R. D. Multiple routes to astrocytic differentiation in the CNS. *J. Neurosci.* **18**, 3620–3629 (1998).
- Yang, J. *et al.* Novel roles of unphosphorylated STAT3 in oncogenesis and transcriptional regulation. *Cancer Res.* **65**, 939–947 (2005).
- Pullen, N. *et al.* Phosphorylation and activation of p70^{S6K} by PDK1. *Science* **279**, 707–710 (1998).
- Alessi, D. R., Kozlowski, M. T., Weng, Q. P., Morrice, N. & Avruch, J. 3-Phosphoinositide-dependent protein kinase 1 (PDK1) phosphorylates and activates the p70 S6 kinase *in vivo* and *in vitro*. *Curr. Biol.* **8**, 69–81 (1998).
- Deak, M., Clifton, A. D., Lucocq, L. M. & Alessi, D. R. Mitogen- and stress-activated protein kinase-1 (MSK1) is directly activated by MAPK and SAPK/p38, and may mediate activation of CREB. *EMBO J.* **17**, 4426–4441 (1998).
- Lavoie, J. N., L'Allemain, G., Brunet, A., Muller, R. & Pouyssegur, J. Cyclin D1 expression is regulated positively by the p42/p44MAPK and negatively by the p38/HOGMAPK pathway. *J. Biol. Chem.* **271**, 20608–20616 (1996).
- Doetsch, F., Petreanu, L., Caille, I., Garcia-Verdugo, J. M. & Alvarez-Buylla, A. EGF converts transit-amplifying neurogenic precursors in the adult brain into multipotent stem cells. *Neuron* **36**, 1021–1034 (2002).
- Berman, D. M. *et al.* Medulloblastoma growth inhibition by hedgehog pathway blockade. *Science* **297**, 1559–1561 (2002).
- Romer, J. & Curran, T. Targeting medulloblastoma: small-molecule inhibitors of the Sonic Hedgehog pathway as potential cancer therapeutics. *Cancer Res.* **65**, 4975–4978 (2005).
- Kamakura, S. *et al.* Hes binding to STAT3 mediates crosstalk between Notch and JAK–STAT signalling. *Nature Cell Biol.* **6**, 547–554 (2004).
- Oishi, K. *et al.* Notch promotes survival of neural precursor cells via mechanisms distinct from those regulating neurogenesis. *Dev. Biol.* **276**, 172–184 (2004).
- Niwa, H., Burdon, T., Chambers, I. & Smith, A. Self-renewal of pluripotent embryonic stem cells is mediated via activation of STAT3. *Genes Dev.* **12**, 2048–2060 (1998).
- Watanabe, S. *et al.* Activation of Akt signaling is sufficient to maintain pluripotency in mouse and primate embryonic stem cells. *Oncogene* **25**, 2697–2707 (2006).
- Pera, M. F. *et al.* Regulation of human embryonic stem cell differentiation by BMP-2 and its antagonist noggin. *J. Cell Sci.* **117**, 1269–1280 (2004).
- Couillard-Despres, S. *et al.* Doublecortin expression levels in adult brain reflect neurogenesis. *Eur. J. Neurosci.* **21**, 1–14 (2005).
- Leker, R. R., Gai, N., Mechoulam, R. & Ovadia, H. Drug-induced hypothermia reduces ischemic damage: effects of the cannabinoid HU-210. *Stroke* **34**, 2000–2006 (2003).
- Zhao, Y., Patzer, A., Gohlke, P., Herdegen, T. & Culman, J. The intracerebral application of the PPAR γ -ligand pioglitazone confers neuroprotection against focal ischaemia in the rat brain. *Eur. J. Neurosci.* **22**, 278–282 (2005).
- Gustafsson, M. V. *et al.* Hypoxia requires notch signaling to maintain the undifferentiated cell state. *Dev. Cell* **9**, 617–628 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank R. Kageyama and T. Kitamura for the Hes3 and STAT3 plasmids. This research was supported in part by the Intramural Research Program of the NIH, NINDS.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.D.G.M. (mckay@codon.nih.gov).

Protein flexibility acclimatizes photosynthetic energy conversion to the ambient temperature

Oksana Shlyk-Kerner^{1*}, Ilan Samish^{1*†}, David Kaftan^{1*†}, Neta Holland^{1†}, P. S. Maruthi Sai^{1†}, Hadar Kless¹ & Avigdor Scherz¹

Adjustment of catalytic activity in response to diverse ambient temperatures is fundamental to life on Earth. A crucial example of this is photosynthesis, where solar energy is converted into electrochemical potential that drives oxygen and biomass generation at temperatures ranging from those of frigid Antarctica to those of scalding hot springs. The energy conversion proceeds by concerted mobilization of electrons and protons on photoexcitation of reaction centre protein complexes^{1–3}. Following physicochemical paradigms, the rates of imperative steps in this process were predicted to increase exponentially with rising temperatures, resulting in different yields of solar energy conversion at the distinct growth temperatures of photosynthetic mesophiles and extremophiles. In contrast, here we show a meticulous adjustment of energy conversion rate, resulting in similar yields from mesophiles and thermophiles. The key molecular players in the temperature adjustment process consist of a cluster of hitherto unrecognized protein cavities and an adjacent packing motif that jointly impart local flexibility crucial to the reaction centre proteins. Mutations within the packing motif of mesophiles that increase the bulkiness of the amino-acid side chains, and thus reduce the size of the cavities, promote thermophilic behaviour. This novel biomechanical mechanism accounts for the slowing of the catalytic reaction above physiological temperatures in contradiction to the classical Arrhenius paradigm. The mechanism provides new guidelines for manipulating the acclimatization of enzymes to the ambient temperatures of diverse habitats. More generally, it reveals novel protein elements that are of potential significance for modulating structure–activity relationships in membrane and globular proteins alike.

The recent structural resolution of the oxygenic photosystem II (PSII) reaction centre^{4,5} highlighted the marked similarity in protein structures and cofactors used to convert solar energy into useful electrochemical potential in very different organisms from diverse ecological niches. This database of fundamentally important membrane-protein family members provides a unique opportunity to search for protein motifs used to regulate adaptation of enzymatic activity to the macroenvironment. The solar energy conversion in reaction centres of non-oxygenic purple bacteria¹ and of PSII in oxygenic organisms² is initiated by means of a multistep, light-induced electron transfer across the photosynthetic membrane⁶ (Fig. 1a). The process ends with the reduction of one quinone molecule (Q_B) by another quinone (Q_A), accompanied by essential reorganization and mobilization of protons³. Extensive studies of this reaction have shown that protein flexibility and a conformational change^{1,7,8} are required to initiate $Q_A^- \rightarrow Q_B(Q_B^-)$ electron transfer,

the rate constant (k) is independent of the redox energies of the quinones¹, and that the reaction should be accelerated at elevated temperatures^{7,9}. Following the classical Arrhenius paradigm, k is expected to increase exponentially with $1/T$ (where T is temperature measured in kelvin (K)) at a rate equal to $\Delta H^\ddagger/R$, where $\Delta H^\ddagger$ is the enthalpy of electron transfer activation. However, we observed that this prediction is violated in intact cells of mesophilic and thermophilic cyanobacteria (Fig. 2a). Here, k levels off at a temperature (T_o) slightly below the growth temperature of the examined organisms. Above T_o (26 °C and 59 °C for mesophilic and thermophilic strains, respectively) the rate becomes temperature independent throughout the entire physiological range and then slightly decreases. In all strains, k reaches a similar maximum value of $\sim 3,000\text{--}3,500\text{ s}^{-1}$. In the thermotolerant organism *Cyanidium caldarium*, it reaches the same value, but at $T_o = 43\text{ °C}$ (data not shown). T_o is substantially lower than the reported denaturation temperature of the reaction centre ($>40\text{ °C}$ in the tested mesophiles¹⁰), ruling denaturation out as the cause of the observed phenomenon (Supplementary Fig. 1). Thus, the photosynthetic organism seems to use novel physicochemical principles to slow down the rate acceleration at elevated temperatures. Remarkably, the activation enthalpies ($\Delta H^\ddagger$) below T_o for the examined mesophile and thermophile are almost identical (Fig. 2a). However, the activation entropy ($\Delta S^\ddagger$) for the thermophilic *Thermosynechococcus elongatus* is more negative by $\sim 1\text{ kcal mol}^{-1}$, indicating that the transition state flexibility for electron transfer in the reaction centre excited state is a key player in the temperature adjustment of the energy conversion rate.

In searching for motifs that account for the observed differences in the activation entropies, we examined the sequences of the two major protein subunits found in all purple bacteria^{1,11} and PSII reaction centres (Supplementary Fig. 2). Rather than searching in proximity to the cofactor-binding sites, we confined our search to the four transmembrane-helices that are involved in holding the electron transfer cofactors (Fig. 1a, b). Two sites, D1-209 and D1-212, were found to show consistent changes between mesophilic, thermotolerant and thermophilic organisms including cyanobacteria, algae and green plants. The sites are positioned in a GXXXG-like sequence motif (where G denotes small residues such as Gly, Ala, Ser, Cys and Thr; Supplementary Fig. 2)¹², found at the points of closest contact between the two major protein subunits, D1 and D2 (refs 4, 5). This motif, and the structurally homologous motif in purple bacteria reaction centres (Supplementary Fig. 2), participates in an inter-subunit hydrogen bonding (ISHB) network (Fig. 1; Supplementary Fig. 3 and Supplementary Table 1). Such motifs in the central part of membrane proteins were suggested to provide local flexibility while

¹Department of Plant Sciences, The Weizmann Institute of Science, Rehovot 76100, Israel. [†]Present addresses: Department of Biochemistry and Molecular Biophysics, School of Medicine and Department of Chemistry, School of Arts and Sciences, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA (I.S.); Laboratory of Nanobiology, Institute of Physical Biology USB, and Institute of Systems Biology and Ecology ASCR, Zamek 136, CZ-37333 Nove Hrad, Czech Republic (D.K.); Department of Biology, Israel Institute of Technology, Haifa 32000, Israel (N.H.); Biologics Division, Biological E. Limited 18/1&3, Azamabad, Hyderabad, Andhra Pradesh 500020, India (P.S.M.S.).

*These authors contributed equally to this work.

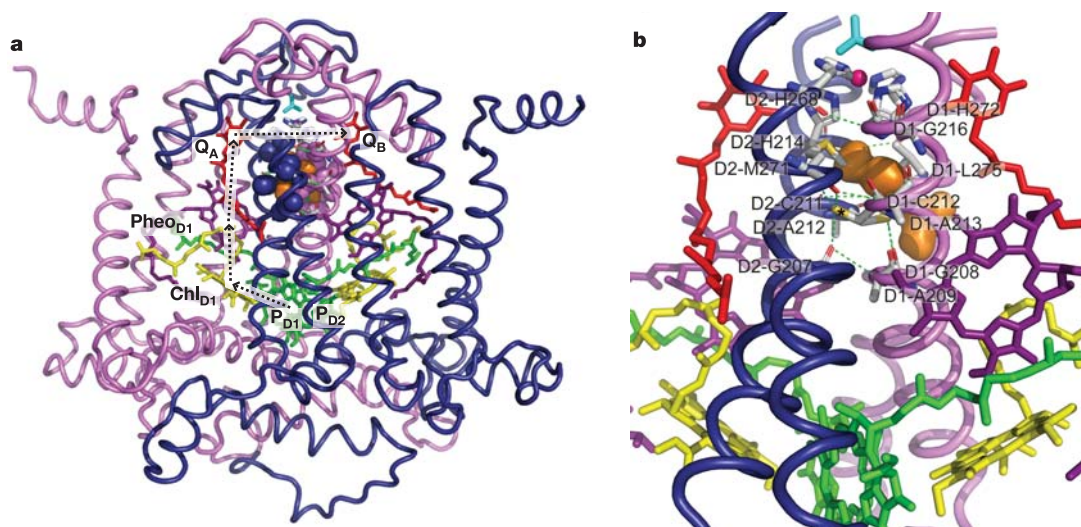


Figure 1 | Conserved structural and functional elements in reaction centres of purple-bacteria and of PSII. **a**, Light induces electron transfer (dashed arrows) from a special pair of (bacterio)chlorophylls (P_{D1} and P_{D2}, green) through an accessory (bacterio)chlorophyll (Chl_{D1}, yellow) and (bacterio)pheophytin (Pheo_{D1}, purple) to quinones (Q_A and Q_B, red) via cofactors held by the D1 (pink) and D2 (blue) subunits. The non-haem iron

(magenta) and the atoms lining the three main cavities (orange) are depicted. **b**, Potential ISHB (dashed lines) between helices of D1 and D2 of the two subunits. Residues (identified with single-letter code) involved in these bonds, for example, D1-C212 (sulphur of cysteine marked with a star) are depicted. For clarity, only the two largest cavities are drawn.

maintaining overall stability^{12,13}. Thus, changes in the amino acids within the identified GXXXG-like motif may result in modification of the local flexibility of the reaction centre and, consequently, in the entropy of activation. Indeed, Ser occupies D1-212 and D1-209 in mesophilic organisms including the cyanobacterium *Synechocystis* 6803; the D1-209 is modified to Ala in the thermotolerant red alga *Cyanidium caldarium*; in thermophilic cyanobacteria Cys or Ala populates D1-212 whereas Ala occupies the D1-209 site (Supplementary Fig. 2).

Two sizable cavities augmented by several small cavities complement the ISHB cluster (Fig. 1a, b; Supplementary Fig. 3 and Supplementary Table 2). The largest cavity (47 Å³) lies at the interface of the two subunits and extends from the D1-212 site towards the histidines that ligate the non-haem iron. A second cavity (32 Å³) occupies the other side of the D1-212 residue and is lined by two flexible side chains that are in common with the largest cavity. Consequently, movements or changes in the size of the D1-212 residue may alter the volume and shape of either cavity, or merge the two cavities into one. In the structurally homologous domain in reaction centres of purple bacteria, a large cavity encompassing over 100 Å³ is found in this region. As demonstrated for bacteriorhodopsin¹⁴, such cavities in the inner core of a protein may facilitate the conformational rearrangement required for enhancing local flexibility during discrete functional steps.

Having two adjacent structural motifs associated with flexibility, we tentatively proposed that replacement of Ser by Cys or Ala within the GXXXG-like motif accounts for the observed changes in local protein flexibility and, subsequently, the temperature adjustment of reaction centre activity. To test our assumption, we used the available structures from thermophilic cyanobacteria^{4,5} to perform *in silico* rotamer-library-based saturated mutagenesis on the D1-212 site, conservatively allowing only the D1-212 to repack. Under such constraints, the volume of the larger cavities was reduced by ~10–20% and the smaller cavities merged into one. This new cavity was evident in only the mutants that had small residues (Gly, Ala, Cys, Ser; Supplementary Table 2). In all other mutants the D1-212 side chain blocked the cavity (Fig. 3a, b; Supplementary Table 2). Dissociation of ISHB between residues of the GXXXG-like motif should increase the localized flexibility provided that side chains are free to undergo reorganization. However, this freedom of motion is

decreased when the packing values¹⁵ of the involved residues are decreased, shifting *T*₀ to higher values. Thus, Ala and Cys (packing value of 0.471 and 0.463, respectively) are expected to correspond to higher *T*₀ values compared with Ser (packing value of 0.484). Unlike the non-bulky residues (termed class I), the bulky residues (termed class II) partly block the intersubunit cavity (Fig. 3a, b; Supplementary Table 2). The latter have the lowest packing values and are expected to elevate *T*₀ to a non-physiological range at which PSII reaction centre denaturation occurs.

We verified the *in silico* predictions directly by introducing mutations in the PSII reaction centre of *Synechocystis* 6803. The D1-S212 residue was replaced by each of the remaining 19 amino acids and the effect on the photosynthetic parameters in intact cells was investigated. Thirteen photoautotrophic mutant strains were isolated (see Supplementary Methods 1). Mutants with aromatic and positively charged residues (all assigned to class II) failed to grow photoautotrophically, possibly owing to impaired reaction centre folding. Mutants containing class I residues (Gly, Cys, Ala, Thr, Asn and Asp) presented similar patterns of Q_A^{•−} → Q_B electron transfer temperature dependence (Fig. 2b). However, *T*₀ was elevated

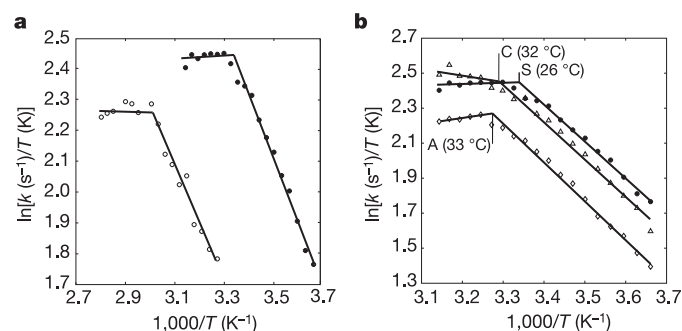


Figure 2 | The Q_A^{•−} → Q_B electron transfer rate in mesophilic and thermophilic cyanobacteria. **a**, Eyring plots for electron transfer in a mesophile (*Synechocystis* 6803, filled circles) and a thermophile (*Thermosynechococcus elongatus*, open circles). **b**, Eyring plots for electron transfer in D1-212 class I mutants: Cys (triangles) and Ala (diamonds) compared with the *Synechocystis* 6803 Ser (filled circles) wild type.

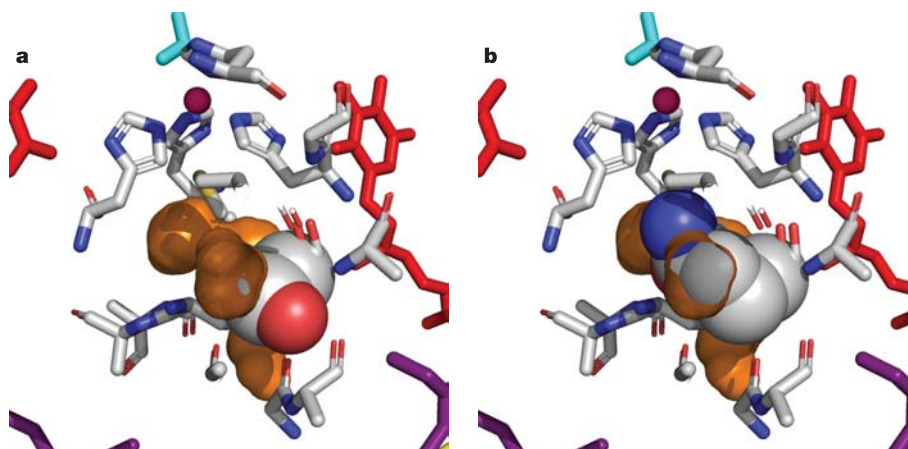


Figure 3 | The effect of D1-212 residues on the PSII reaction centre cavity in *T. elongatus* structure⁴. **a**, A class I residue Ser (space-fill representation of side chain) maintains the cavity open. **b**, A class II residue Gln abolishes one

of the cavities and lines the largest cavity, which is otherwise lined by flexible side chains.

by $\geq 6^\circ\text{C}$ and $\geq 7^\circ\text{C}$ in the D1-S212C and D1-S212A mutants, respectively. The rate constant reached similar maximum values ($\sim 3\text{--}4 \times 10^3 \text{ s}^{-1}$) in the different class I mutants at and above their respective T_0 values (Fig. 2b). Those photoautotrophic mutants with class II residues (Pro, Val, Ile, Leu, Glu and Gln) presented substantially lower values for the electron transfer rate constant, and their k values increased in a monotonic and monophasic fashion with increasing temperature over most of the physiological range (Fig. 4a). The activation enthalpy $\Delta H^\ddagger$ was similar for all class I residues but $\Delta S^\ddagger$ at room temperature was more negative by 0.15 and $0.3 \text{ kcal mol}^{-1}$ for the Cys- and Ala-substituted strains, respectively (Figs 2b and 4b), depicting the significance of D1-212 in modulating the local flexibility of reaction centres and T_0 . For class II mutants, $\Delta H^\ddagger$ was usually similar to wild-type and class I mutants (below T_0), but $\Delta S^\ddagger$ was more negative by up to $\sim 1.8 \text{ kcal mol}^{-1}$ at room temperature (Fig. 4b), thus indicating decreased local flexibility of the transition state for $Q_A^- \rightarrow Q_B$ electron transfer in class II mutants.

Mutagenesis of the D1-S209 residue resulted in only eight photoautotrophic mutants, possibly because substitution with residues of very low packing values impaired the functional assembly of PSII reaction centre: six of class I (Gly, Ala, Thr, Asn and Asp) and only two from class II (Val and Pro). All photoautotrophic mutants showed levelling off of the $Q_A^- \rightarrow Q_B$ electron transfer rate at elevated temperatures (data not shown). The T_0 temperature of the D1-S209A

mutant was elevated by $\geq 7^\circ\text{C}$, while maintaining a maximal rate constant similar to that of the wild type. This finding supports the suggested significance of the D1-209 residue in modifying the local flexibility of the PSII reaction centre and thereby the value of T_0 .

Collectively, the identified structural motifs (ISHB in a GXXXG-like sequence motif, and adjacent cavities) in reaction centres of purple bacteria and of PSII, seem to function as key modulators in controlling the activation entropy of $Q_A^- \rightarrow Q_B$ electron transfer and, subsequently, the value of T_0 .

Non-Arrhenius behaviour of enzymatic rate constants was thought to reflect inactivation and denaturation of the examined enzymes. However, T_0 seems independent of PSII reaction centre denaturation, which occurred at substantially higher temperatures (Supplementary Fig. 1). In searching for an alternative explanation for the temperature dependence of k , we looked at the non-Arrhenius behaviour in hydrogen-transfer enzymes¹⁶. Islands of local flexibility within these globular proteins were suggested to enable hydrogen/proton tunnelling at elevated temperatures^{16–18} making catalysis temperature-independent within the physiological range. Notably, the flexible domains could be fairly distal to the catalytic site¹⁹. These experimental observations are in agreement with recent theoretical developments^{16,20}, and apply also to proton-coupled electron transfer²¹. As redistribution of negative charges in the reaction centre protein as a result of the $Q_A^- \rightarrow Q_B$ electron transfer should be compensated with proton reorganization²², then proton tunnelling may account for the levelling of the electron transfer rate constant reported here.

In conclusion, we showed that changes in locally flexible domains in membrane proteins provide the means for enzyme adaptation to the ambient temperature. We further described and validated protein motifs that modulate this adaptation independent of denaturation processes. We believe that similar motifs and mechanisms may account for the acclimation of other systems, including globular proteins^{23,24}, to significantly different habitats.

METHODS

In silico analysis. Rotamer-based virtual mutagenesis was conducted using the SCWRL²⁵ program to mutate D1-212 to all possible residues. Hydrogens were added with REDUCE²⁶, and intersubunit hydrogen bonds²⁷ were screened using an in-house program²⁸. Intra-protein cavities were analysed using VOLBL²⁹. Figures were drawn using PyMol version 0.99 (DeLano Scientific). See also Supplementary Methods 2.

Cell growth and mutagenesis. *Synechocystis* sp. PCC 6803 strains were grown on BG-11 agar plates at $50 \mu\text{mol (photons)} \text{ m}^{-2} \text{ s}^{-1}$ of white light at 30°C in the presence of 5 mM glucose and then transferred into a stirred or air bubbled

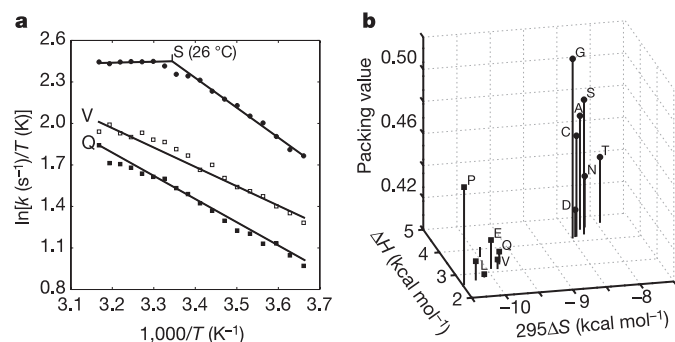


Figure 4 | Temperature dependence and activation parameters of the $Q_A^- \rightarrow Q_B$ electron transfer in D1-212 mutants of *Synechocystis* 6803. **a**, Eyring plots of D1-S212 (filled circles), D1-V212 (open squares) and D1-Q212 (filled squares). **b**, Correlation between the activation entropy, enthalpy for the $Q_A^- \rightarrow Q_B$ electron transfer and the packing values of residues at the D1-212 site. Class I residues (filled circles) are distinct from the class II residues (filled squares).

250 ml BG-11 solution for further growth in the absence of glucose. Saturating mutagenesis of the *psbAII* gene encoding the D1 protein was performed as described in see Supplementary Methods 1.

Fluorescence measurement. Chlorophyll fluorescence decay was measured in a FL-100 double-modulation fluorometer equipped with a TR 2000 thermoregulator (P.S. Instruments). The normalized decay kinetics were fitted by least-squares numerical analysis to a three-exponential function and the highest rate constant was assigned to the electron transfer from Q_A^- to Q_B , as described³⁰. Importantly, fluorescence analysis as used here allowed for discrimination between active and inactive PSII reaction centres¹⁰. The inactive component contributes to the longest lifetime component and becomes apparent when detecting the PSII reaction centre activity at temperatures well above the physiological range¹⁰.

Received 21 March; accepted 2 June 2006.

Published online 23 July 2006.

- Okamura, M. Y., Paddock, M. L., Graige, M. S. & Feher, G. Proton and electron transfer in bacterial reaction centers. *Biochim. Biophys. Acta* **1458**, 148–163 (2000).
- Iwata, S. & Barber, J. Structure of photosystem II and molecular architecture of the oxygen-evolving centre. *Curr. Opin. Struct. Biol.* **14**, 447–453 (2004).
- Wraight, C. A. Proton and electron transfer in the acceptor quinone complex of photosynthetic reaction centers from *Rhodobacter sphaeroides*. *Front. Biosci.* **9**, 309–337 (2004).
- Ferreira, K. N., Iverson, T. M., Maghlaoui, K., Barber, J. & Iwata, S. Architecture of the photosynthetic oxygen-evolving center. *Science* **303**, 1831–1838 (2004).
- Loll, B., Kern, J., Saenger, W., Zouni, A. & Biesiadka, J. Towards complete cofactor arrangement in the 3.0 Å resolution structure of photosystem II. *Nature* **438**, 1040–1044 (2005).
- Moser, C. C. et al. in *Advances in Protein Chemistry 63: Membrane Proteins* (ed. Rees, D. C.) 71–109 (Academic, San Diego, 2003).
- Li, J., Gilroy, D., Tiede, D. M. & Gunner, M. R. Kinetic phases in the electron transfer from $P^+Q_A^-Q_B$ to $P^+Q_A^-Q_B^-$ and the associated processes in *Rhodobacter sphaeroides* R-26 reaction centers. *Biochemistry* **37**, 2818–2829 (1998).
- Garbers, A., Reifarh, F., Kurreck, J., Renger, G. & Parak, F. Correlation between protein flexibility and electron transfer from Q_A^- to Q_B in PSII membrane fragments from spinach. *Biochemistry* **37**, 11399–11404 (1998).
- Xu, Q. & Gunner, M. R. Exploring the energy profile of the Q_A^- to Q_B electron transfer reaction in bacterial photosynthetic reaction centers: pH dependence of the conformational gating step. *Biochemistry* **41**, 2694–2701 (2002).
- Joshi, M. & Fragata, M. Heat-induced changes in the photochemical centres and the protein secondary structures of photosystem II studied by variable fluorescence and difference FT-IR spectroscopy. *Z. Naturforsch.* **54**, 35–43 (1999).
- Deisenhofer, J., Epp, O., Sinning, I. & Michel, H. Crystallographic refinement at 2.3 Å resolution and refined model of the photosynthetic reaction centre from *Rhodospseudomonas viridis*. *J. Mol. Biol.* **246**, 429–457 (1995).
- Senes, A., Engel, D. E. & DeGrado, W. F. Folding of helical membrane proteins: the role of polar, GxxxG-like and proline motifs. *Curr. Opin. Struct. Biol.* **14**, 465–479 (2004).
- Curran, A. R. & Engelman, D. M. Sequence motifs, polar interactions and conformational changes in helical membrane proteins. *Curr. Opin. Struct. Biol.* **13**, 412–417 (2003).
- Friedman, R., Nachliel, E. & Gutman, M. The role of small intraprotein cavities in the catalytic cycle of bacteriorhodopsin. *Biophys. J.* **85**, 886–896 (2003).
- Eilers, M., Shekar, S., Shieh, T., Smith, S. & Fleming, P. Internal packing of helical membrane proteins. *Proc. Natl Acad. Sci. USA* **97**, 5796–5801 (2000).
- Knapp, M. J., Rickert, K. & Klinman, J. P. Temperature-dependent isotope effects in soybean lipoxygenase-1: Correlating hydrogen tunneling with protein dynamics. *J. Am. Chem. Soc.* **124**, 3865–3874 (2002).
- Kohen, A., Cannio, R., Bartolucci, S. & Klinman, J. P. Enzyme dynamics and hydrogen tunnelling in a thermophilic alcohol dehydrogenase. *Nature* **399**, 496–499 (1999).
- Liang, Z. X., Lee, T., Resing, K. A., Ahn, N. G. & Klinman, J. P. Thermal-activated protein mobility and its correlation with catalysis in thermophilic alcohol dehydrogenase. *Proc. Natl Acad. Sci. USA* **101**, 9556–9561 (2004).
- Wong, K. F., Selzer, T., Benkovic, S. J. & Hammes-Schiffer, S. Impact of distal mutations on the network of coupled motions correlated to hydride transfer in dihydrofolate reductase. *Proc. Natl Acad. Sci. USA* **102**, 6807–6812 (2005).
- Kohen, A. & Klinman, J. P. Enzyme catalysis: Beyond classical paradigms. *Acc. Chem. Res.* **31**, 397–404 (1998).
- Hammes-Schiffer, S. & Iordanova, N. Theoretical studies of proton-coupled electron transfer reactions. *Biochim. Biophys. Acta* **1655**, 29–36 (2004).
- Remy, A. & Gerwert, K. Coupling of light-induced electron transfer to proton uptake in photosynthesis. *Nature Struct. Biol.* **10**, 637–644 (2003).
- Somero, G. N. Protein adaptations to temperature and pressure: complementary roles of adaptive changes in amino-acid sequence and internal milieu. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* **136**, 577–591 (2003).
- Závadoszky, P., Kardos, J., Svingor, A. & Petsko, G. A. Adjustment of conformational flexibility is a key event in the thermal adaptation of proteins. *Proc. Natl Acad. Sci. USA* **95**, 7406–7411 (1998).
- Bower, M. J., Cohen, F. E. & Dunbrack, R. L. Jr. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. *J. Mol. Biol.* **267**, 1268–1282 (1997).
- Word, J. M. et al. Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. *J. Mol. Biol.* **285**, 1711–1733 (1999).
- Senes, A., Ubarretxena-Belandia, I. & Engelman, D. M. The $C\alpha-H\cdots O$ hydrogen bond: A determinant of stability and specificity in transmembrane helix interactions. *Proc. Natl Acad. Sci. USA* **98**, 9056–9061 (2001).
- Goldberg, E. *Membrane Protein Assembly: The Role of Backbone Mediated Interhelical Hydrogen Bonds*. Master's thesis, Weizmann Institute of Science (2004).
- Liang, J., Edelsbrunner, H., Fu, P., Sudhakar, P. V. & Subramaniam, S. Analytical shape computation of macromolecules: II. Inaccessible cavities in proteins. *Proteins* **33**, 18–29 (1998).
- Nedbal, L., Trtílek, M. & Kaftan, D. Flash fluorescence induction: a novel method to study regulation of photosystem II. *J. Photochem. Photobiol. B* **48**, 154–157 (1999).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank B. Loll, J. Biesiadka, and W. Saenger for sending us their PS II reaction centre coordinates before publication; M. Edelman for critical reading of the manuscript; S. Petrokovski for assistance in data mining sequences; and W. L. DeLano for assistance in applying new PyMol methods. This research was supported by a Sonderforschungsbereich grant and the Willstatter-Avron-Minerva Foundation for Photosynthesis. D.K. was also supported by the Ministry of Education, Youth, and Sports of the Czech Republic; the Academy of Sciences of the Czech Republic; and by the NIH NBCR. A.S. is the incumbent of the Yadelle and Robert Sklare Professorial Chair for Biochemistry.

Author Contributions A.S. carried out project planning and supervision. O.S. designed and performed mutagenesis. O.S. and D.K. performed experimental work, mainly at the WIS and partly at USB and ACSR. O.S., D.K. and I.S. performed comparative sequence analysis. I.S. performed the *in silico* study. O.S., D.K., P.S.M.S., I.S. and A.S. carried out experimental data analysis. H.K. and N.H. helped to establish the mutagenesis program. The study was done in partial fulfilment of the PhD theses of O.S. and I.S.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.S. (Avigdor.Scherz@weizmann.ac.il).

Structure of the catalytic domain of the hepatitis C virus NS2-3 protease

Ivo C. Lorenz^{1*}, Joseph Marcotrigiano^{1*}, Thomas G. Dentzer¹ & Charles M. Rice¹

Hepatitis C virus is a major global health problem affecting an estimated 170 million people worldwide¹. Chronic infection is common and can lead to cirrhosis and liver cancer. There is no vaccine available and current therapies have met with limited success². The viral RNA genome encodes a polyprotein that includes two proteases essential for virus replication^{3,4}. The NS2-3 protease mediates a single cleavage at the NS2/NS3 junction, whereas the NS3-4A protease cleaves at four downstream sites in the polyprotein. NS3-4A is characterized as a serine protease with a chymotrypsin-like fold^{5,6}, but the enzymatic mechanism of the NS2-3 protease remains unresolved⁷⁻⁹. Here we report the crystal structure of the catalytic domain of the NS2-3 protease at 2.3 Å resolution. The structure reveals a dimeric cysteine protease with two composite active sites. For each active site, the catalytic histidine and glutamate residues are contributed by one monomer, and the nucleophilic cysteine by the other. The carboxy-terminal residues remain coordinated in the two active sites, predicting an inactive post-cleavage form. Proteolysis through formation of a composite active site occurs in the context of the viral polyprotein expressed in mammalian cells. These features offer unexpected insights into polyprotein processing by hepatitis C virus and new opportunities for antiviral drug design.

Crystallization of the catalytic domain of the hepatitis C virus (HCV) NS2-3 protease (NS2^{Pro}, consisting of residues 94–217 of NS2; Fig. 1a) using native and selenomethionine-containing protein yielded two crystal forms with the same space group (*P*₂₁, Supplementary Table 1). The asymmetric units of the native and selenomethionine-containing protein contained twelve and six NS2^{Pro} molecules organized into six and three tightly packed dimers, respectively (Supplementary Fig. 1).

The NS2^{Pro} monomer consists of two subdomains connected by an extended linker (Fig. 1b and Supplementary Fig. 2a). The amino-terminal subdomain contains two antiparallel α -helices (H1 and H2) followed by several turns and loops that contact both H1 and H2. The polypeptide chain continues into an extended region before entering a four-stranded, antiparallel β -sheet in the C-terminal subdomain. The last β -strand continues to the C terminus of NS2. Figure 1c, d represents views of the NS2^{Pro} dimer, which resembles a 'butterfly' with two-fold symmetry along the vertical axis (Fig. 1c). The N-terminal subdomain of one molecule interacts with the C-terminal subdomain of the other molecule and *vice versa*. The two extended linkers cross over in the middle of each molecule and each contribute a β -strand (b1) to the antiparallel β -sheet in the C-terminal subdomain of the other molecule. The N termini of the two monomers lie relatively close to each other, whereas the solvent-exposed C termini are positioned on opposite sides of the molecule.

Critical residues for NS2-3 proteolytic activity^{7,8}, His 143 and Glu 163, are located in the loop region following helix H2 in the N-terminal subdomain, whereas another critical residue, Cys 184, lies at the end of the linker arm in the b1–b2 loop of the C-terminal subdomain. At the dimer interface, the histidine and glutamate residues from one monomer are close to the cysteine from the other chain (Fig. 2a). The arrangement of these three residues is suggestive of a composite cysteine protease active site (Fig. 2b and Supplementary Fig. 2b).

The NS2^{Pro} structure represents a novel protein fold (DALI server¹⁰ Z scores of less than 3.0). However, superimposing His 143, Glu 163 and Cys 184 of NS2^{Pro} with the active sites from cysteine proteases such as papain¹¹ (Fig. 2c) and poliovirus 3C protease¹² (Fig. 2d) demonstrated a similar spatial distribution. The orientation of the catalytic cysteine residues from NS2^{Pro} and 3C^{Pro} is similar to the catalytic serine residues of Sindbis virus capsid¹³ (Fig. 2e) and the cellular protease subtilisin¹⁴ (Fig. 2f). Thus, like poliovirus 3C^{Pro}, HCV NS2-3 is a cysteine protease with a serine protease active site geometry¹⁵. To the best of our knowledge, NS2^{Pro} represents the first example of a cysteine or serine protease that forms a dimer containing a pair of composite active sites. However, these features are reminiscent of retroviral aspartic proteases, which consist of dimers with a single active site at the dimerization interface¹⁶. Other proteases such as caspases¹⁷ require dimerization for activity, but they do not contain composite active sites.

Interestingly, Pro 164, which is entirely conserved in all HCV sequences, has a *cis*-peptide conformation (Fig. 2b and Supplementary Fig. 2c). Pro 164 may bend the peptide backbone of the catalytic Glu 163 to establish the correct geometry of the glutamate side chain for catalysis. In addition, the *cis*-proline may contribute to dimer stabilization as the linker connecting the two subdomains follows Pro 164.

The backbone carboxylic acid of the C-terminal residue of NS2^{Pro}, Leu 217, remains coordinated in the active site via contacts to the side chains of His 143 and Cys 184, and to the backbone nitrogen of Cys 184 (Fig. 2b). Comparison of the NS2^{Pro} structure with protease-inhibitor complexes indicates that the C terminus of NS2 may exert an inhibitory function after cleavage⁸. Sindbis virus capsid protein mediates a single autoproteolytic cleavage at its C terminus, which remains bound to the active site inhibiting further catalysis¹³. The C-terminal residues of Sindbis virus capsid (Trp 264) and HCV NS2 (Leu 217) are in the same orientation relative to the catalytic triad (Fig. 2e). Comparing the structure of NS2^{Pro} with subtilisin bound to the inhibitor Eglin-C¹⁴ demonstrates that the C terminus of NS2^{Pro} occupies a position equivalent to the inhibitor (Fig. 2f). We propose a model in which each NS2^{Pro} molecule catalyses a single NS2-3 cleavage event, allowing tightly regulated processing. However, we cannot rule out the possibility that NS2 mediates proteolysis of other

¹Laboratory of Virology and Infectious Disease, Center for the Study of Hepatitis C, The Rockefeller University, 1230 York Avenue, New York, New York 10021, USA.

*These authors contributed equally to this work.

viral or cellular proteins if the C-terminal β -strand (b5) is displaced from the active site.

Processing at the NS2/NS3 junction requires the NS3 protease domain and is stimulated by the addition of exogenous zinc^{7,8,18,19}. However, because NS2 contains a complete cysteine-protease active site with the C terminus positioned for catalysis, the function of NS3 remains undetermined. The crystal structure of NS2^{Pro} represents the post-cleavage form, which may differ from the NS2-3 precursor. NS3 may interact with the highly conserved surface surrounding the NS2 active site (Supplementary Fig. 3), contributing to a functional

catalytic environment and correct positioning of the scissile bond. The backbone nitrogen of Cys 184 contacts the carboxylic acid of Leu 217 and may serve as part of the oxyanion hole to stabilize the transition state during catalysis. A residue in uncleaved NS2-3 (either a backbone nitrogen of NS2 oriented differently in the pre-cleavage form or a residue within the NS3 serine protease domain) may also contribute to the oxyanion hole. The zinc requirement may be due to its structural function in NS3 (refs 5, 6) rather than a role in NS2-3 catalysis. Thus, limiting zinc could indirectly inhibit NS2-3 cleavage by affecting NS3 folding²⁰. Consistent with this idea, the NS2-3 protease is inhibited by mutations in zinc-coordinating residues of NS3 (refs 7, 8).

Molecular surface analysis and biochemical data support the NS2^{Pro} dimer model. NS2^{Pro} shows a high degree of amino-acid sequence conservation at the interface between the two monomers (Supplementary Fig. 3). The cleavage rate of purified NS2-3 is concentration dependent, indicating that the active form of the protease is oligomeric¹⁹. Analytical ultracentrifugation of NS2^{Pro} yielded a single, monodisperse species with a molecular weight of 39 kDa that most likely corresponds to a dimer with bound detergent (data not shown). Moreover, cross-linking of NS2^{Pro} in solution with

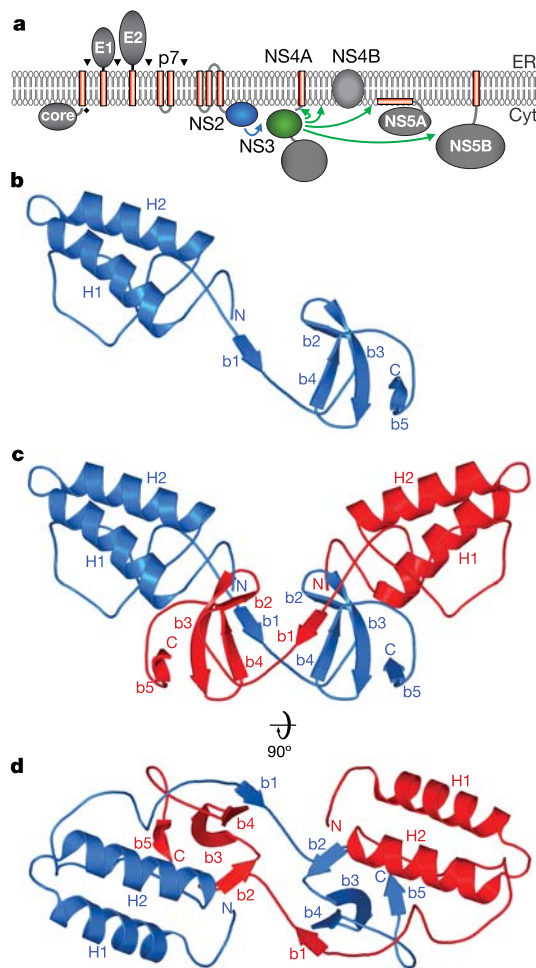


Figure 1 | Processing of the HCV polyprotein and architecture of NS2^{Pro}.

a, Membrane association of the HCV polyprotein, which contains a core, envelope proteins E1 and E2, p7, and nonstructural (NS) proteins NS2, NS3, NS4A, NS4B, NS5A and NS5B. Cleavage in the structural region and at the N terminus of NS2 occurs by action of the host signal peptidase (filled arrowheads) and host signal peptide peptidase (diamond). The NS2-3 protease, which consists of residues 94–217 of NS2 (blue) and residues 1–181 of NS3 (green), cleaves at the NS2/NS3 junction (blue arrow). All cleavages downstream of NS3 are mediated by the NS3-4A protease (green arrows). The model shown here depicts NS2 with three N-terminal transmembrane segments followed by the protease domain on the cytoplasmic face of the endoplasmic reticulum (ER) membrane¹⁹. Alternatively, the NS2 C terminus may be localized to the ER lumen^{28,29}. Cyt, cytosol. **b**, Ribbon diagram showing the NS2 monomer. The N and C termini, and secondary structure elements (helices H1 and H2; β -strands b1–b5), are labelled. **c**, Ribbon diagram showing the NS2 dimer with one monomer in blue, the other in red. The N and C termini, and secondary structure elements of each monomer, are indicated. **d**, Ribbon diagram of the NS2 dimer viewed perpendicular to helices H1 and H2. This view is a 90° rotation around the horizontal axis in **c**.

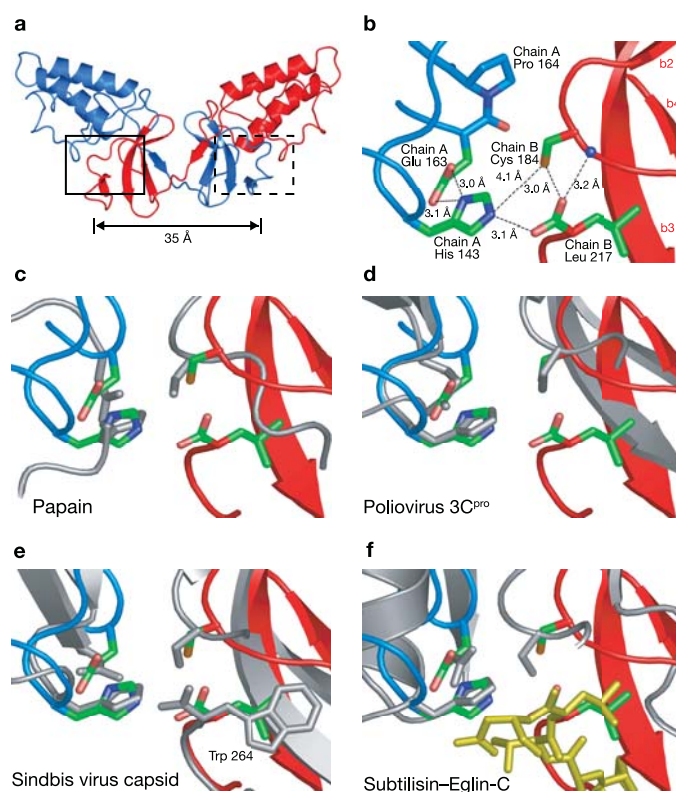


Figure 2 | The active site of NS2 and comparison with other proteases.

a, Location of the two active sites in the NS2 dimer (boxed regions). The solid-lined box represents the active site displayed in **b–f**. The distance between the two active sites is indicated. **b**, The NS2 active site. Residues His 143, Glu 163, Pro 164, Cys 184 and Leu 217 are shown as stick drawings. The active site is composed of His 143 and Glu 163 from one molecule of the dimer (chain A, drawn in blue), and Cys 184 from the other molecule (chain B, drawn in red). The C-terminal residue, Leu 217, originates from the same chain as Cys 184. Dashed lines indicate contacts between selected residues. The length of each contact is provided. **c–f**, Superimposition of His 143, Glu 163 and Cys 184 of NS2 with similar catalytic sites of the proteases papain (**c**), poliovirus 3C protease (**d**), Sindbis virus capsid (**e**) and subtilisin bound to the inhibitor Eglin-C (coloured yellow) (**f**). The orientation and colouring of the NS2 active site is identical to **b**, with the other proteases shown in grey. In **e**, the C-terminal residue Trp 264 of Sindbis virus capsid is labelled.

disuccinimidyl suberate (DSS) led to the identification of a dimeric species (Supplementary Fig. 4).

A series of experiments in mammalian cells was designed to test whether NS2^{pro} can form dimers with a functional composite active site *in vivo*. HCV full-length polyproteins containing either a H143A or a C184A mutation in the NS2 active site are defective in NS2-3 processing^{7,8}. However, if a composite active site can form, co-expression of the two mutant polyproteins should result in partial NS2-3 cleavage (Fig. 3a). Indeed, when HCV polyproteins with NS2 containing either a H143A or C184A mutation were co-expressed, NS2 and NS3 cleavage products were detected (Fig. 3b), indicating the formation of a functional composite active site. Moreover, it is possible to predict from the crystal structure which mutant polypeptide in the mixing experiment is cleaved, because the C-terminal Leu 217 and the catalytic Cys 184 originate from the same chain. Thus,

mixing of the two mutants should lead to cleavage of NS2-3(H143A), whereas NS2-3(C184A) is predicted to remain unprocessed. By expressing NS2-3 proteins with a Flag or haemagglutinin (HA) tag fused to the N terminus of NS2, it is possible to distinguish Flag-NS2 from HA-NS2 by immunoprecipitation using epitope-specific antibodies and by different electrophoretic mobilities of the various polypeptides (Fig. 3c). Mixing of Flag-NS2-3(H143A) with HA-NS2-3(C184A) resulted in cleavage of Flag-NS2-3, whereas HA-NS2-3 remained unprocessed (Fig. 3c, second lane from the right). When HA-NS2-3(H143A) was mixed with Flag-NS2-3(C184A), only HA-NS2-3 was cleaved (Fig. 3c, first lane from the right). Mixing wild-type Flag-NS2-3 with double-mutant HA-NS2-3(H143A/C184A) or *vice versa* yielded only cleaved wild-type NS2 (Fig. 3d), whereas the double-mutant polypeptide remained unprocessed because neither of the composite active sites is functional when a wild-type and a double-mutant NS2-3 dimerize. Finally, when cells are co-transfected with wild-type Flag-NS2-3 and HA-NS2-3 and lysed with a mild detergent, Flag-NS2 and HA-NS2 can be co-precipitated using either an anti-Flag or an anti-HA antibody (Fig. 3e). These data strongly support the NS2^{pro} crystal structure and prove that NS2 can form dimers with composite functional active sites.

Our data change the current view of HCV polyprotein processing and raise interesting regulatory possibilities. Previously, NS2-3 cleavage was thought to occur as a unimolecular reaction in *cis*. The apparent requirement for dimerization to form an active NS2-3 protease suggests that NS2-3 cleavage and therefore formation of the active RNA replicase may be dependent on the concentration of NS2. Thus, a requirement for NS2-3 dimerization and subsequent proteolytic processing could delay the initiation of RNA replication, which may allow the virus to accumulate sufficient amounts of active NS3-4A protease to antagonize the induction of type I interferons by the host²¹.

The cleaved NS2 dimer and higher-order oligomers may have additional roles in the viral life cycle, such as a function in membrane-associated virus assembly²². The solvent-accessible surface of the NS2^{pro} dimer, coloured according to electrostatic potential (Fig. 4a, b), showed a high proportion of neutral and basic regions. The surface of the molecule near helices H1 and H2 is mainly hydrophobic, with basic residues lying underneath. The crystal structure of NS2^{pro} contained several molecules of *n*-octyl- β -glucoside and *n*-decyl- β -maltoside interacting with those two helices (Fig. 4d, e and Supplementary Fig. 2d). We propose a model in which NS2^{pro} interacts peripherally with cellular membranes (Fig. 4c, f). The N termini of two NS2^{pro} monomers would lie

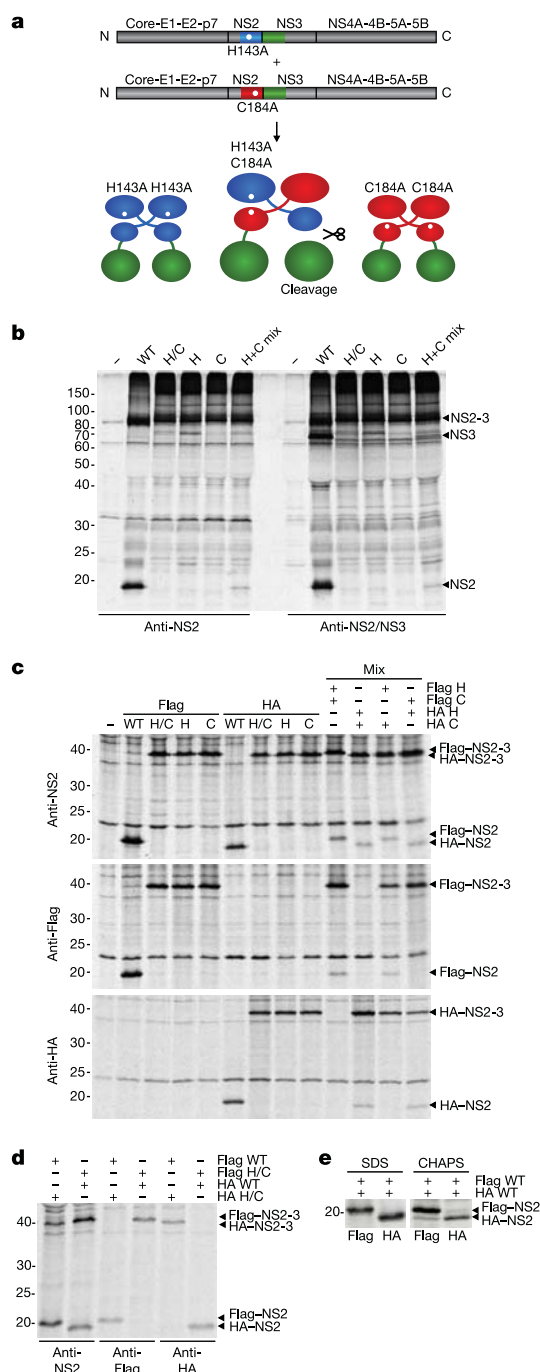


Figure 3 | Dimerization of NS2 and formation of a composite active site in mammalian cells. **a**, Mixing of two HCV polyproteins, each containing a point mutation in the NS2 active site (H143A or C184A; shown in blue and red, respectively) leads to the formation of NS2-3 mutant homo- and hetero-dimers. Dimerization between NS2(H143A) and NS2(C184A) yields one defective and one functional active site, resulting in partial NS2-3 cleavage. The protease domain of NS3 is shown in green. **b**, Mixing experiment in Huh-7.5 cells of HCV full-length polyproteins containing a wild-type (WT), single- (H or C) or double-mutant (H/C) NS2 active site, followed by metabolic labelling, cell lysis and immunoprecipitation using anti-NS2 or anti-NS2-3 antibodies. **c**, Co-transfection of U2OS cells with Flag- and HA-tagged NS2-3 constructs (consisting of residues 19–217 of NS2 and 1–181 of NS3) containing a wild-type or mutant NS2 active site as indicated on top of the lanes, followed by metabolic labelling, cell lysis and immunoprecipitation using anti-NS2, anti-Flag or anti-HA antibodies. **d**, Mixing experiment of Flag–NS2-3 and HA–NS2-3 containing a wild-type or double-mutant NS2 active site, as described above. **e**, Co-immunoprecipitation using anti-Flag or anti-HA antibodies of co-transfected Flag–NS2-3 and HA–NS2-3, lysed either in SDS or CHAPS buffer. Positions of the NS2, NS3 and NS2-3 proteins are shown by arrowheads on the right. Molecular weight markers are indicated on the left. H, H143A; C, C184A; H/C, H143A/C184A.

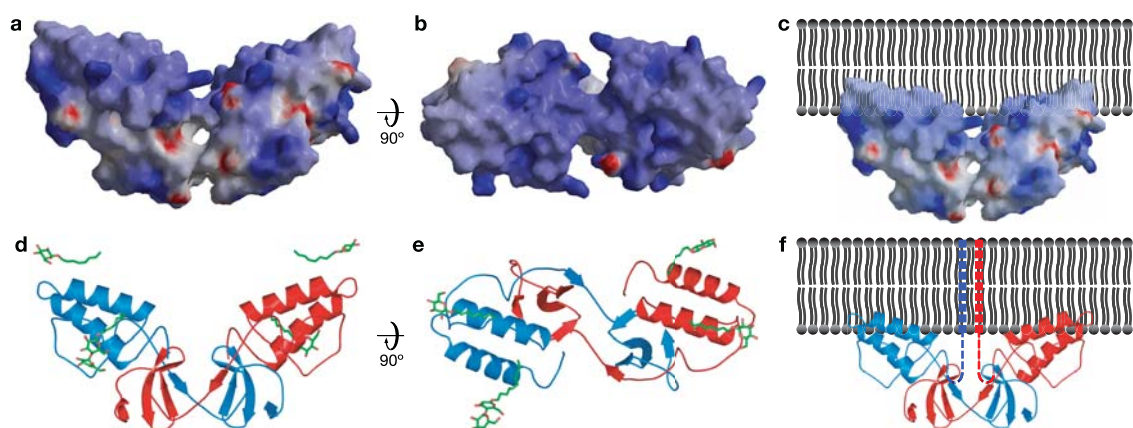


Figure 4 | Model for membrane association of NS2. **a**, Solvent-accessible surface of the NS2 dimer coloured according to electrostatic potential: acidic (red), neutral (white) and basic (blue). **b**, Electrostatic surface of the NS2 dimer rotated 90° around the horizontal axis shown in **a**. **c**, Electrostatic surface of the NS2 dimer inserted peripherally into a cellular membrane. **d**, Ribbon diagram of the NS2 dimer oriented as in **a**, with three *n*-octyl- β -glucoside molecules and one *n*-decyl- β -maltoside molecule bound to the N-terminal subdomains. **e**, Ribbon diagram of the NS2 dimer oriented as in

b, showing the four bound detergent molecules. **f**, Ribbon diagram of the NS2 dimer positioned relative to the membrane. In this model, the hydrophobic part of helix H2 from each subunit is peripherally inserted into the lipid bilayer, with basic amino acids on the side of the helix involved in neutralizing the charge of polar lipid head groups in the membrane. The N-terminal transmembrane segments of NS2 (shown as dotted lines) would extend into the membrane.

close to the membrane (Fig. 4f). In the full-length protein, dimerization of NS2 may cause the N-terminal transmembrane domains of two monomers to form a 'bundle' of transmembrane segments that may serve an important function during virion morphogenesis.

The NS2^{pro} structure presented here will allow further studies to elucidate the role of NS2 dimerization and other functions of NS2 in the viral life cycle. In addition, the structure establishes a foundation for the design of small-molecule inhibitors directed against the well-defined active site cleft and other conserved features of the protein.

METHODS

See Supplementary Information for detailed methods.

Protein preparation. NS2^{pro} (NS2 residues 94–217), a truncated form of NS2 shown to be proteolytically active in the context of an NS2-3 precursor that includes the NS3 protease domain^{18,19}, was expressed in *Escherichia coli*. Lysis and subsequent purification of the protein by immobilized-metal affinity chromatography, ion exchange chromatography and gel filtration was done in the presence of detergent. The final concentration on a cation exchange column yielded highly pure protein at 6–9 mg ml⁻¹.

Crystal growth and freezing. Crystals of NS2^{pro} were grown using hanging-drop vapour diffusion at 4°C. The well contained a solution of 0.1 M Tris, pH 8.5, 0.8 M ammonium acetate, 0.25 M lithium chloride and 12% PEG 3350. Crystals were frozen in well solution supplemented with 5.4 mM decyl maltoside using stepwise addition of glycerol to a final concentration of 25%.

Data collection and structure determination. Data were collected at beamlines X9A and X29 at the Brookhaven National Laboratory's National Synchrotron Light Source. Phases were calculated from selenomethionine-containing protein by multiwavelength anomalous diffraction (MAD). After data indexing and scaling with DENZO/SCALEPACK²³, 19 of the 24 selenium sites were found using SnB²⁴. An interpretable electron density map was obtained using MLPHARE²⁵, followed by density modification and phase combination using SOLOMON and DM²⁵. Several rounds of iterative model building and refinement were done using O²⁶ and CNS²⁷. The 2.9 Å resolution structure was used as a search model to obtain phases for a native data set at 2.28 Å resolution using molecular replacement. The final model contained 176 solvent molecules, 12 detergent molecules, and 12 molecules of NS2^{pro}. A summary of the refinement statistics is shown in Supplementary Table 1.

Expression in mammalian cells. NS2 containing wild-type or mutant active site residues was expressed in mammalian cells in the context of a full-length viral polyprotein or as NS2-3 precursor with an N-terminal Flag or HA tag. Cells metabolically labelled with ³⁵S were lysed and immunoprecipitation of NS2 performed using antibodies against NS2 or the Flag and HA tags. Proteins were separated by SDS-polyacrylamide gel electrophoresis and visualized by autoradiography.

Received 8 May; accepted 9 June 2006.

Published online 23 July 2006.

- Alter, H. J. & Seeff, L. B. Recovery, persistence, and sequelae in hepatitis C virus infection: a perspective on long-term outcome. *Semin. Liver Dis.* **20**, 17–35 (2000).
- Hoofnagle, J. H. Course and outcome of hepatitis C. *Hepatology* **36**, S21–S29 (2002).
- Bartenschlager, R., Frese, M. & Pietschmann, T. Novel insights into hepatitis C virus replication and persistence. *Adv. Virus Res.* **63**, 71–180 (2004).
- Lindenbach, B. D. & Rice, C. M. in *Fields Virology* (eds Knipe, D. M. et al.) 991–1041 (Lippincott Williams & Wilkins, Philadelphia, Pennsylvania, 2001).
- Love, R. A. et al. The crystal structure of hepatitis C virus NS3 proteinase reveals a trypsin-like fold and a structural zinc binding site. *Cell* **87**, 331–342 (1996).
- Kim, J. L. et al. Crystal structure of the hepatitis C virus NS3 protease domain complexed with a synthetic NS4A cofactor peptide. *Cell* **87**, 343–355 (1996).
- Grakoui, A., McCourt, D. W., Wychowski, C., Feinstone, S. M. & Rice, C. M. A second hepatitis C virus-encoded proteinase. *Proc. Natl Acad. Sci. USA* **90**, 10583–10587 (1993).
- Hijikata, M. et al. Two distinct proteinase activities required for the processing of a putative nonstructural precursor protein of hepatitis C virus. *J. Virol.* **67**, 4665–4675 (1993).
- Wu, Z., Yao, N., Le, H. V. & Weber, P. C. Mechanism of autoproteolysis at the NS2–NS3 junction of the hepatitis C virus polyprotein. *Trends Biochem. Sci.* **23**, 92–94 (1998).
- Holm, L. & Sander, C. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123–138 (1993).
- Pickersgill, R. W., Harris, G. W. & Garman, E. Structure of monoclinic papain at 1.60 Å resolution. *Acta Crystallogr. B* **48**, 49–66 (1992).
- Mosimann, S. C., Cherney, M. M., Sia, S., Plotch, S. & James, M. N. Refined X-ray crystallographic structure of the poliovirus 3C gene product. *J. Mol. Biol.* **273**, 1032–1047 (1997).
- Tong, L., Wengler, G. & Rossmann, M. G. Refined structure of Sindbis virus core protein and comparison with other chymotrypsin-like serine proteinase structures. *J. Mol. Biol.* **230**, 228–247 (1993).
- McPhalen, C. A. & James, M. N. Structural comparison of two serine proteinase-protein inhibitor complexes: eglin-c-subtilisin Carlsberg and Cl-2-subtilisin Novo. *Biochemistry* **27**, 6582–6598 (1988).
- Malcolm, B. A. The picornaviral 3C proteinases: cysteine nucleophiles in serine proteinase folds. *Protein Sci.* **4**, 1439–1445 (1995).
- Wlodawer, A. & Gustchina, A. Structural and biochemical studies of retroviral proteases. *Biochim. Biophys. Acta* **1477**, 16–34 (2000).
- Boatright, K. M. & Salvesen, G. S. Mechanisms of caspase activation. *Curr. Opin. Cell Biol.* **15**, 725–731 (2003).
- Thibeault, D., Maurice, R., Pilote, L., Lamarre, D. & Pause, A. *In vitro* characterization of a purified NS2/3 protease variant of hepatitis C virus. *J. Biol. Chem.* **276**, 46678–46684 (2001).
- Pallaoro, M. et al. Characterization of the hepatitis C virus NS2/3 processing reaction by using a purified precursor protein. *J. Virol.* **75**, 9939–9946 (2001).

20. Urbani, A. *et al.* The metal binding site of the hepatitis C virus NS3 protease. A spectroscopic investigation. *J. Biol. Chem.* **273**, 18760–18769 (1998).
21. Gale, M. Jr & Foy, E. M. Evasion of intracellular host defence by hepatitis C virus. *Nature* **436**, 939–945 (2005).
22. Pietschmann, T. *et al.* Construction and characterization of infectious intragenotypic and intergenotypic hepatitis C virus chimeras. *Proc. Natl Acad. Sci. USA* **103**, 7408–7413 (2006).
23. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
24. Weeks, C. M. & Miller, R. The design and implementation of SnB v2.0. *J. Appl. Crystallogr.* **32**, 120–124 (1999).
25. Collaborative Computational Project No. 4, The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
26. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
27. Brunger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
28. Santolini, E., Pacini, L., Fipaldini, C., Migliaccio, G. & Monica, N. The NS2 protein of hepatitis C virus is a transmembrane polypeptide. *J. Virol.* **69**, 7461–7471 (1995).
29. Yamaga, A. K. & Ou, J. H. Membrane topology of the hepatitis C virus NS2 protein. *J. Biol. Chem.* **277**, 33228–33234 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We would like to thank R. MacKinnon, S. Darst and

H. Mueller for the use of X-ray diffractometers, related equipment and software. We appreciate access to beamlines X9A and X29 of the National Synchrotron Light Source (NSLS) at the Brookhaven National Laboratory and acknowledge the assistance of the NSLS staff. We wish to thank T. Tellinghuisen for assistance with data collection. D. Jeruzalmi provided the program `msf_similarity_to_pdb` and `LOOSENGRASP`. We would like to thank S. Burley, S. Darst, L. Dustin, M. Evans, C. Jones, C. Murray, B. Lindenbach, G. Randall and T. Tellinghuisen for input on the manuscript, and K. Kuo and S. You for help with artwork. I.C.L. was supported by fellowships from the Swiss National Science Foundation, the Roche Research Foundation and the Swiss Society for Medical-Biological Grants. J.M. was supported as a Merck Fellow of the Life Sciences Research Foundation. Additional financial support for this work came from grants from the National Institutes of Health and the Greenberg Medical Research Institute (C.M.R.).

Author Contributions I.C.L., J.M. and C.M.R. conceived the experiments. I.C.L. generated all reagents, materials, proteins and crystals with assistance from J.M. All data collection and processing was done by J.M. and I.C.L. Model building and refinement were performed by I.C.L. and J.M. Cross-linking and mammalian expression studies were carried out by I.C.L. and T.G.D. The manuscript was written by I.C.L., J.M. and C.M.R.

Author Information The atomic coordinates for this structure have been deposited in the Protein Data Bank under accession code 2HDO. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to C.M.R. (ricec@rockefeller.edu) or J.M. (marcotj@rockefeller.edu).

LETTERS

DNA overwinds when stretched

Jeff Gore^{1†}, Zev Bryant^{2,4‡}, Marcelo Nöllmann², Mai U. Le², Nicholas R. Cozzarelli^{2‡} & Carlos Bustamante^{1–4}

DNA is often modelled as an isotropic rod^{1–4}, but its chiral structure suggests the possible importance of anisotropic mechanical properties, including coupling between twisting and stretching degrees of freedom. Simple physical intuition predicts that DNA should unwind under tension, as it is pulled towards a denatured structure^{4–8}. We used rotor bead tracking to directly measure twist–stretch coupling in single DNA molecules. Here we show that for small distortions, contrary to intuition, DNA overwinds under tension, reaching a maximum twist at a tension of ~ 30 pN. As tension is increased above this critical value, the DNA begins to unwind. The observed twist–stretch coupling predicts that DNA should also lengthen when overwound under constant tension, an effect that we quantitatively confirm. We present a simple model that explains these unusual mechanical properties, and also suggests a possible origin for the anomalously large torsional rigidity of DNA. Our results have implications for the action of DNA-binding proteins that must stretch and twist DNA to compensate for variability in the lengths of their binding sites^{9–11}. The requisite coupled DNA distortions are favoured by the intrinsic mechanical properties of the double helix reported here.

Many cellular proteins bend or wrap DNA upon binding, loop DNA to make contact with two non-adjacent binding sites, or twist DNA during translocation along the double helix¹². The energetics of these distortions are governed by the mechanical properties of DNA, which have been investigated using a variety of bulk¹ and single-molecule techniques^{2,3}. For small deformations, DNA in physiological buffer is modelled as an isotropic rod with bending rigidity $B = 230 \pm 20$ pN nm², twist rigidity $C = 460 \pm 20$ pN nm², and stretch modulus $S = 1,100 \pm 200$ pN (refs 3, 13–16; and Z.B., J.G., N.R.C. and C.B., manuscript in preparation).

A fourth mechanical parameter is allowed in the linear theory of a deformable rod^{5,6,17}: the twist–stretch coupling g , which specifies how the twist of the helix changes when the molecule is stretched. At forces sufficient to suppress bending fluctuations, the energy of a stretched and twisted DNA molecule may be written as^{5,6,17}:

$$E_{\text{DNA}} = \frac{1}{2} \frac{C}{L} \theta^2 + g \theta \frac{x}{L} + \frac{1}{2} \frac{S}{L} x^2$$

where L is the contour length at zero force, x is the distance that the DNA is stretched beyond its contour length L , and θ is the angle through which the DNA is twisted from its unperturbed equilibrium value.

Interpolation between the B-form helix and denatured or over-stretched forms of DNA suggests that g should be positive^{5,6}, so that DNA unwinds as it is stretched (in an analogous process to the unwinding of DNA at elevated temperatures¹⁸). Previous fits^{5–7} to experimental single-molecule data^{4,8,14} have indeed yielded positive g values ($g = 200 \pm 100$ pN nm). However, an analysis of the distribution of base pair step parameters in atomic structures of

DNA:protein complexes showed a weak positive correlation between twist and rise, implying a negative g value⁹. All-atom simulations have likewise suggested that twist and rise may be positively correlated for small distortions^{10,19}. Finally, in contrast to the overstretching transition^{8,13,14} (in which DNA unwinds as it extends), the B–A transition involves a slight unwinding coupled to compression of the DNA helix²⁰. Conclusive determination of the sign and magnitude of g requires direct measurement in isolated DNA molecules.

To measure the twist–stretch coupling of DNA, we used the rotor bead tracking technique^{13,15,21}, in which a submicrometre ‘rotor’ bead is attached to the middle of a stretched DNA molecule, immediately below a free swivel consisting of an engineered single strand nick (Fig. 1a, b). Tension is applied to the molecule using magnetic tweezers^{2,4}. Changes in the rotor bead angle reflect changes in the twist of the lower DNA segment.

Under fixed tension, the rotor bead fluctuated around a mean angle as a result of thermal noise (Fig. 1c). As the molecule was stretched by increasing the magnetic force, the mean angle of the rotor bead increased. A $\sim 1\%$ stretching of DNA led to an increase in the twist of $\sim 0.1\%$ (Fig. 1d). Analysis of this data yielded a twist–stretch coupling constant $g = -90 \pm 20$ pN nm ($N = 4$ molecules), opposite in sign to most previous estimates^{5–7} but consistent with twist–rise correlations in crystal structures⁹ and molecular simulations^{10,19}. Our measurement of g was robust to alterations in the length and sequence of the torsionally constrained DNA segment (Fig. 1d).

For small deformations, we have shown that DNA overwinds when stretched. However, in the limit of high forces, DNA must eventually unwind as the backbone is pulled straight. To test for sign reversal of g at elevated tensions, we monitored the twist of DNA molecules while gradually increasing the magnetic force. We found that the twist of DNA increases until the tension reaches a critical value $F_c \approx 30$ pN, beyond which the DNA begins to unwind (Fig. 1e).

The negative twist–stretch coupling g observed at lower tensions implies that DNA should lengthen when overwound (Fig. 2a, b). To predict the expected magnitude of this effect, we write the total energy of the DNA/magnetic bead system as

$$E_T = \frac{1}{2} \frac{C}{L} \theta^2 + g \theta \frac{x}{L} + \frac{1}{2} \frac{S}{L} x^2 - xF$$

and minimize it with respect to the stretching distance x , holding the twist θ and force F constant ($(\partial E_T / \partial x)_{\theta, F} = 0$). This analysis yields the DNA extension, x^* , that minimizes the total energy given some imposed twist value^{5,6}:

$$x^* = \frac{L}{S} \left(F - \frac{g\theta}{L} \right) \Rightarrow \frac{\partial x^*}{\partial \theta} = -\frac{g}{S}$$

Given our measured $g = -90 \pm 20$ pN nm, we expect the DNA molecule to lengthen by $\Delta x = 0.5 \pm 0.1$ nm for each rotation imposed in the overwinding direction.

¹Department of Physics, ²Department of Molecular and Cell Biology, ³Howard Hughes Medical Institute, University of California, Berkeley, California 94720, USA. ⁴Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. [†]Present addresses: Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA (J.G.); Department of Biochemistry, Stanford University School of Medicine, Stanford, California 94305, USA (Z.B.).

[‡]Deceased.

To measure changes in extension upon imposed overwinding, we employed the magnetic tweezers assay introduced in refs 4 and 22, in which a single constrained DNA molecule is stretched between a coverslip and a magnetic bead (Fig. 2a, b). The twist in the molecule can be controlled via rotation of the magnets, and extension is monitored by measuring the focal depth of the bead using image analysis. Overwinding caused the DNA to extend by 0.5 nm per turn, in quantitative agreement with our prediction (Fig. 2c, d).

What could be the physical origin of the negative twist–stretch coupling of DNA? A model helix with a fixed backbone length and fixed radius must necessarily unwind as it is stretched. However, if the radius of the helix is allowed to shrink as the helix is stretched, then unwinding is no longer guaranteed. In simulations of DNA stretching using all-atom potentials¹⁰, a reduction in the radius of the double helix has been seen concurrent with stretching and overwinding. We constructed a simple ‘toy’ model in which the radius of the helix was allowed to vary as the molecule was stretched, and asked whether this model could capture the twist–stretch coupling and other mechanical properties of DNA.

The model consists of an elastic rod with a stiff helical ‘wire’ (analogous to the sugar-phosphate backbone) affixed to the outside surface (Fig. 3a). The inner rod is constructed from a material with a Poisson’s ratio $\nu = 0.5$, so that it conserves volume under stress²³. As this system is stretched, the inner rod decreases in diameter. When the model is stretched without constraining twist, changes in helicity arise from the tendency of the stiff outer ‘wire’ to resist changes in contour length (Fig. 3b). The inner rod alone has stretch modulus $S_r = \pi R_r^2 Y_r$, bending rigidity $B_r = \pi R_r^4 Y_r / 4$ and torsional rigidity

$C_r = B_r / (1 + \nu)$, where Y_r is the Young’s modulus of the rod material and R_r is the rod’s radius^{24,25}.

For small deformations, the complete toy model has the following effective mechanical parameters (see Supplementary Information):

$$B_{\text{eff}} \approx B_r$$

$$S_{\text{eff}} = S_r + S_h \csc \alpha (\sin^2 \alpha - \nu \cos^2 \alpha)^2 \approx S_r$$

$$C_{\text{eff}} = C_r + R_r^2 S_h \sin \alpha \cos^2 \alpha$$

$$g = R_r (\sin^2 \alpha - \nu \cos^2 \alpha) S_h$$

where S_h is the stretch modulus of the outer wire, and $\alpha = \arctan(3.4 \text{ nm} / 2\pi R_r)$ is the helix angle (Fig. 3a). Thus, the presence of the outer wire does not change the bending rigidity or stretch modulus appreciably, but it does stiffen the molecule to torsion and also generates a non-zero twist–stretch coupling, g . With the three free parameters fitted to $R_r = 0.924 \text{ nm}$ (close to the crystallographic radius of DNA), $S_h = 965 \text{ pN}$, and $Y_r = 0.393 \text{ GPa}$ (similar to estimated Y values for DNA and within the range of measured Y values for bulk polymeric materials^{14,24,26}), we obtain the correct experimentally measured values for the mechanical parameters of DNA: $B_{\text{eff}} = 225 \text{ pN nm}^2$, $C_{\text{eff}} = 460 \text{ pN nm}^2$, $g = -90 \text{ pN nm}$ and $S_{\text{eff}} = 1,081 \text{ pN}$.

Although construction of this toy DNA model was motivated by the discovery of negative twist–stretch coupling, it also provides a possible explanation for the anomalously large torsional rigidity of DNA. For an isotropic rod, the torsional rigidity C must be smaller than the bending rigidity B (see equation for C_r above) unless the rod

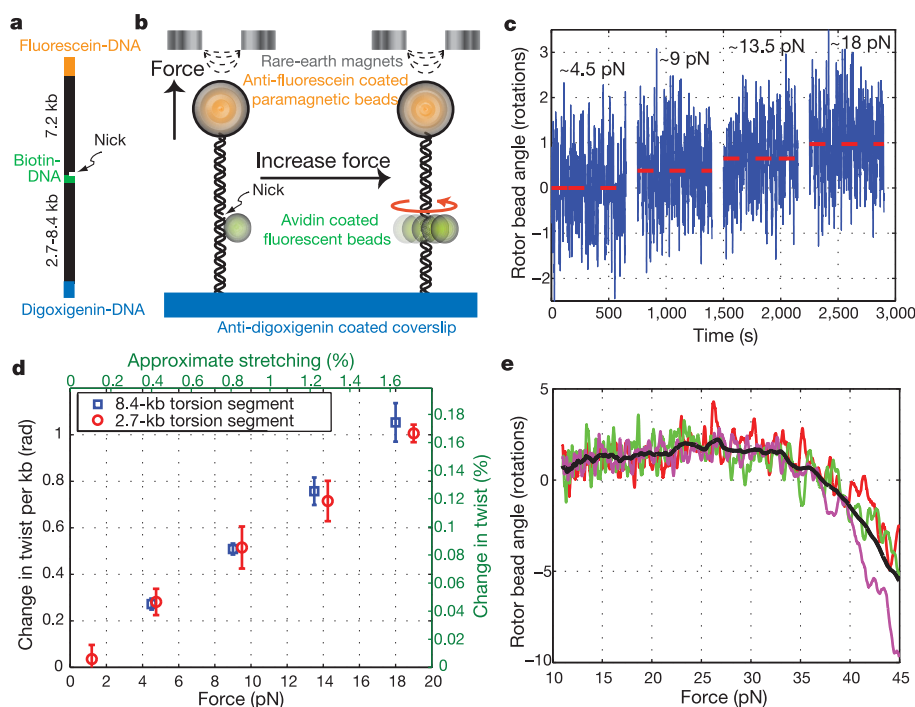


Figure 1 | DNA overwinds when stretched. **a**, The molecular construct for rotor bead tracking experiments contains three distinct attachment sites and a site-specific nick, which acts as a swivel^{13,15,21}. **b**, Molecule/bead assemblies were constructed in parallel in a flow chamber, and assayed with an inverted microscope equipped with permanent magnets²¹. Each molecule was stretched between the glass coverslip and a magnetic bead, while a fluorescent avidin-coated rotor bead was attached to the central biotinylated patch. Tension in the DNA was controlled by raising or lowering the magnets, and changes in twist were observed by tracking the rotation of the fluorescent bead. **c**, When the DNA molecule is held at a fixed force, the rotor bead angle (blue trace) fluctuates around a mean (red dashed lines). As tension is increased, the mean rotor bead angle increases, reflecting

overwinding of the DNA. **d**, The overwinding scales linearly with applied tension and with the length of the torque-bearing DNA segment. Plotted data (mean $\pm$ s.e.m.) correspond to an 8.4-kb segment (blue squares) and a 2.7-kb segment (red circles). **e**, Rotor bead angle versus force during experiments in which the DNA tension was gradually increased by lowering the magnets (8.4-kb segment). Three different experiments are shown in colour; they were averaged and smoothed to obtain the solid black trace. The DNA overwinds until the tension reaches $\sim 30 \text{ pN}$; as the tension is increased above this critical value, the molecule begins to unwind. Equivalent results have also been obtained with DNA constructs containing the 2.7-kb torque-bearing segment (not shown).

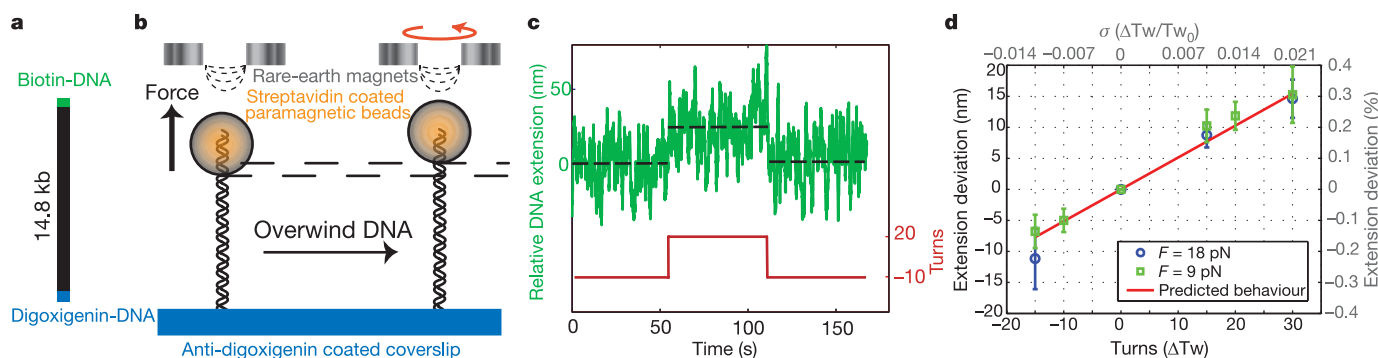


Figure 2 | DNA extends when overwound under constant tension.

a, b, Rotating magnets⁴ were used to introduce torque into a single 14.8-kb DNA tether. **c,** Raw data trace demonstrating that overwinding the DNA molecule at constant tension (9 pN) caused the DNA to extend. **d,** Relative extension as a function of the number of excess turns at 9 pN (green squares)

and 18 pN (blue circles). Each data point shows the mean $\pm$ s.e.m. for a minimum of three molecules. The red line is the predicted behaviour based upon our prior determination of the twist–stretch coupling, $g = -90$ pN nm. Alternative axes show the percentage increase in length induced by a fractional increase in twist, σ .

is constructed from a material with a negative Poisson's ratio (having the unlikely property that the radius increases as the rod is stretched). It has therefore long been puzzling that most measurements of C have been larger than B for DNA²⁷. However, unlike the isotropic rod model, our toy model displays the correct values of B and C without resorting to exotic material properties. The outer helix stiffens the system to torsion because any twisting of the model DNA requires stretching or compression of the outer helical 'wire'.

In the toy model, negative twist–stretch coupling occurs only for shallow helix angles below a critical angle $\alpha_c = \arctan(\sqrt{\nu}) \approx 0.62$ rad. The geometry of B-form DNA lies just within the regime of negative coupling $\alpha = \arctan(3.4 \text{ nm}/2\pi R_r) \approx 0.53 \text{ rad} < \alpha_c$. Although the model is only intended to capture the behaviour of DNA for small distortions, it does predict a sign reversal of twist–stretch coupling upon stretching, reminiscent of the behaviour we observe at high tensions (Fig. 1e).

The mechanical properties of DNA have been studied at the single-molecule level for over a decade; why is it that the surprising

twist–stretch coupling has not been observed experimentally in the past? One study⁷ obtained the incorrect sign for the coupling by analysing relatively noisy magnetic-tweezers data⁴ that were generated to study DNA buckling, an effect with a much larger signal than the twist–stretch coupling. In the final phases of preparing this manuscript, we learned that Croquette and co-workers had performed a new set of magnetic-tweezers measurements, focusing on changes in extension of DNA under small changes in twist. They find that overwinding causes DNA to extend²⁸, in agreement with the results presented in Fig. 2.

As discussed in ref. 10, the anomalous twist–stretch coupling of DNA has important implications for plasticity in site recognition by DNA-binding proteins. Consider the effect of a base-pair insertion or deletion in the binding site for a protein. Even if the change does not directly involve a DNA:protein contact, the protein must overcome a geometric mismatch in both extension and twist in order to bind the DNA. And yet, proteins are able to recognize binding sites with variable sequence lengths; this can be achieved by simultaneously stretching and overwinding (or compressing and underwinding) the DNA. An extreme example occurs in the 146-base-pair (bp) nucleosome core particle structure¹¹. Equivalent binding sites in the two halves of the nucleosome are occupied by 13 bp of DNA in one case and only 12 bp in the other. For the 12-bp sequence to fit into its binding site, it must be stretched to the same length as the 13-bp sequence, and also overwound to give the same total twist as the 13-bp sequence. This large distortion is outside the range of negative twist–stretch coupling we have observed for random sequences, but single-basepair insertions or deletions in longer sequences will lead to more modest distortions. In addition, larger effects might be facilitated by sequence-dependent mechanics: coupling between twist and rise may be particularly strong for certain base-pair steps⁹, and thus certain sequences may be optimized for coupled twisting and stretching. This hypothesis may be tested by future single-molecule studies of the sequence dependence of twist–stretch coupling.

METHODS

The rotor bead tracking experiments were performed in an inverted epifluorescence microscope (Zeiss Axiovert 100A) equipped with permanent magnets mounted on a motorized translation stage, as described previously^{15,21}. The applied force was estimated to within 20% by bright-field imaging of the transverse fluctuations of the magnetic bead^{4,15}. The DNA construct was prepared by serial ligation as described previously¹³, with the torque-bearing DNA segment replaced by the 8.4-kilobase (kb) *BglII-SalI* fragment of pSV8¹³ or (where noted) the 2.7-kb *BamHI-SalI* fragment of pUC18. Each molecule was stretched between a glass coverslip coated with anti-digoxigenin (Roche) and a 2.8- μm magnetic bead (DynaM-280) coated with anti-fluorescein (Molecular Probes), while a fluorescent avidin-coated rotor bead (SpheroTech VFP-0552-5, nominal diameter 0.46 μm) was attached to the central biotinylated patch.

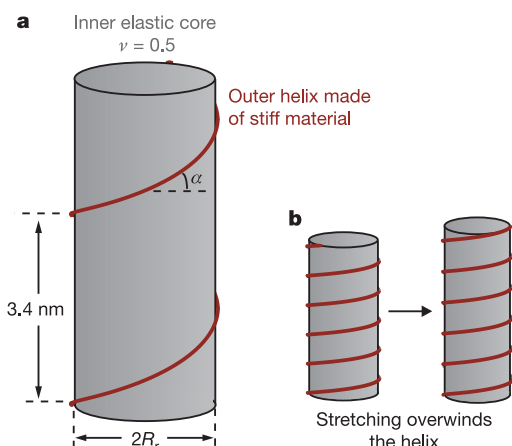


Figure 3 | Toy mechanical model of DNA. **a,** We model DNA as an elastic rod (grey) wrapped helically by a stiff wire (red). The inner core of radius R_r is assumed to have a Poisson's ratio $\nu = 0.5$. The outer helix is affixed to the inner rod helically with a pitch of 3.4 nm, and contributes to the overall mechanical properties because it resists stretching and compression. The outer helix increases the torsional rigidity and yields a twist–stretch coupling that depends upon the helix angle, α . **b,** Stretching generates an overwinding of the helix because the inner rod decreases in diameter as it is stretched. The outer helix is then able to wrap a larger number of times over the length of the molecule. In this figure, a shallow helix angle was used in order to exaggerate the overwinding effect seen with DNA.

Movies of the fluorescent rotor beads were recorded at 100 Hz on an electron-multiplying CCD camera (Andor iXon DU-860E-CS0-#BV). Experiments were performed at room temperature ($23 \pm 2^\circ\text{C}$) in 40 mM Tris HCl, pH 8.0, 100 mM NaCl, 1 mM EDTA, 0.2% sodium azide, 0.2% Tween-20, $50 \mu\text{g ml}^{-1}$ BSA. Replacing EDTA with 10 mM MgCl_2 led to similar degrees of overwinding with tension.

The effect of imposed twist (Fig. 2) was studied in a purpose-built microscope (S. Hong, D. Humphries, M. D. Stone, C.B. and N.R.C., manuscript in preparation) equipped with high-powered magnets and a piezoelectric objective positioner (Physikinstrumente). Changes in extension of the DNA were measured by tracking the focal depth of the streptavidin-coated magnetic bead (DynaM-280 or MyOne) using image analysis of ring patterns^{4,29}. The effect of drift was minimized by alternating between successive twist values (as in Fig. 2c). The DNA tether for these experiments was the 14.8-kb *Bam*HI-*Sal*I fragment from pPIA2-6³⁰.

Received 13 March; accepted 8 June 2006.

Published online 12 July 2006.

- Hagerman, P. J. Flexibility of DNA. *Annu. Rev. Biophys. Biophys. Chem.* **17**, 265–286 (1988).
- Smith, S. B., Finzi, L. & Bustamante, C. Direct mechanical measurements of the elasticity of single DNA molecules by using magnetic beads. *Science* **258**, 1122–1126 (1992).
- Bustamante, C., Smith, S. B., Liphardt, J. & Smith, D. Single-molecule studies of DNA mechanics. *Curr. Opin. Struct. Biol.* **10**, 279–285 (2000).
- Strick, T. R., Allemand, J. F., Bensimon, D., Bensimon, A. & Croquette, V. The elasticity of a single supercoiled DNA molecule. *Science* **271**, 1835–1837 (1996).
- Marko, J. F. Stretching must twist DNA. *Europhys. Lett.* **38**, 183–188 (1997).
- Kamian, R., Lubensky, T., Nelson, P. & O'Hern, C. Direct determination of DNA twist-stretch coupling. *Europhys. Lett.* **38**, 237–242 (1997).
- Moroz, J. D. & Nelson, P. Entropic elasticity of twist-storing polymers. *Macromolecules* **31**, 6333–6347 (1998).
- Cluzel, P. et al. DNA: an extensible molecule. *Science* **271**, 792–794 (1996).
- Olson, W. K., Gorin, A. A., Lu, X. J., Hock, L. M. & Zhurkin, V. B. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl Acad. Sci. USA* **95**, 11163–11168 (1998).
- Kosikov, K. M., Gorin, A. A., Zhurkin, V. B. & Olson, W. K. DNA stretching and compression: large-scale simulations of double helical structures. *J. Mol. Biol.* **289**, 1301–1326 (1999).
- Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251–260 (1997).
- Travers, A. A. & Thompson, J. M. T. An introduction to the mechanics of DNA. *Phil. Trans. R. Soc. Lond. A* **362**, 1265–1279 (2004).
- Bryant, Z. et al. Structural transitions and elasticity from torque measurements on DNA. *Nature* **424**, 338–341 (2003).
- Smith, S. B., Cui, Y. & Bustamante, C. Overstretching B-DNA: the elastic response of individual double-stranded and single-stranded DNA molecules. *Science* **271**, 795–799 (1996).
- Gore, J. *Single-Molecule Studies of DNA Twist Mechanics and Gyrase Mechanochemistry*. Thesis, Univ. California, Berkeley, (2005).
- Wang, M. D., Yin, H., Landick, R., Gelles, J. & Block, S. M. Stretching DNA with optical tweezers. *Biophys. J.* **72**, 1335–1346 (1997).
- Nelson, P. *Biological Physics* (Freeman, New York, 2003).
- Depew, D. E. & Wang, J. C. Conformational fluctuations of DNA helix. *Proc. Natl Acad. Sci. USA* **72**, 4275–4279 (1975).
- Lankas, F., Spomer, J., Langowski, J. & Cheatham, T. E. DNA basepair step deformability inferred from molecular dynamics simulations. *Biophys. J.* **85**, 2872–2883 (2003).
- Wahl, M. C. & Sundaralingam, M. Crystal structures of A-DNA duplexes. *Biopolymers* **44**, 45–63 (1997).
- Gore, J. et al. Mechanochemical analysis of DNA gyrase using rotor bead tracking. *Nature* **439**, 100–104 (2006).
- Strick, T. R., Allemand, J. F., Bensimon, D. & Croquette, V. Behavior of supercoiled DNA. *Biophys. J.* **74**, 2016–2028 (1998).
- Feynman, R. *Feynman Lectures on Physics* (Addison Wesley Longman, Reading, Massachusetts, 1970).
- Hogan, M. E. & Austin, R. H. Importance of DNA stiffness in protein-DNA binding specificity. *Nature* **329**, 263–266 (1987).
- Hegner, M., Smith, S. B. & Bustamante, C. Polymerization and mechanical properties of single RecA-DNA filaments. *Proc. Natl Acad. Sci. USA* **96**, 10109–10114 (1999).
- Howard, J. *Mechanics of Motor Proteins and the Cytoskeleton* (Sinauer Associates, Sunderland, Massachusetts, 2001).
- Baumann, C. G., Smith, S. B., Bloomfield, V. A. & Bustamante, C. Ionic effects on the elasticity of single DNA molecules. *Proc. Natl Acad. Sci. USA* **94**, 6185–6190 (1997).
- Lionnet, T., Joubaud, S., Lavery, R., Bensimon, D. & Croquette, V. Wringing out DNA. *Phys. Rev. Lett.* **96**, 178102 (2006).
- Gosse, C. & Croquette, V. Magnetic tweezers: Micromanipulation and force measurement at the molecular level. *Biophys. J.* **82**, 3314–3329 (2002).
- Davenport, R. J., Wu, G. J., Landick, R. & Bustamante, C. Single-molecule study of transcriptional pausing and arrest by *E. coli* RNA polymerase. *Science* **287**, 2497–2500 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We dedicate this work to our friend and colleague N.R. Cozzarelli, who passed away during completion of this research. We thank S. Hong, D. Humphries and M. D. Stone for help with the experiments described in Fig. 2, and S. Smith and A. Spakowitz for discussions. This work was supported by an NIH grant to C.B., the Fannie and John Hertz Foundation (J.G.), and the US Department of Energy.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to C.B. (carlos@alice.berkeley.edu).

CORRIGENDUM

doi:10.1038/nature05141

Structure of the Sec13/31 COPII coat cage

Scott M. Stagg, Cemal Gürkan, Douglas M. Fowler, Paul LaPointe,
Ted R. Foss, Clinton S. Potter, Bridget Carragher & William E. Balch

Nature 439, 234–238 (2006)

Details of the cryo-electron microscopy data set pertaining to the structure of Sec 13/31 reported in this Letter have now been deposited in the Electron Microscopy Database at the European Bioinformatics Institute under the accession number 1232 (http://www.ebi.ac.uk/msd-srv/emsearch/atlas/1232_downloads.html).

naturejobs

**THE CAREERS
MAGAZINE FOR
SCIENTISTS**

Who would have thought that a government body would be encouraging researchers to ask it for extra money? It sounds unusual, but that is exactly what Research Councils UK, the body that represents the combined grant-giving agencies in Britain, is doing. In a report issued this month, the organization laments that its scheme for topping up stipends to help keep scientists in research has been hugely underused.

The report, available at www.rcuk.ac.uk/rcd/enhanced.asp, notes that although money isn't the primary motivator for many young scientists, real-world financial concerns are influencing the direction of their careers. These pressures become greater as young scientists move through their training. A survey conducted for the report found that the minimum contract-research salary of £20,000 (US\$38,000) is "increasingly uncompetitive", especially when compared with PhD stipends. The survey results suggest that a stipend of £30,000 would make the positions more attractive.

This information is not new, and the research councils have already implemented 'enhancements' that can be used to boost salaries and stipends in sectors that find it particularly difficult to retain or recruit staff. But in a survey conducted for the report, only 16% of 275 principal investigators questioned were aware that the scheme existed.

The report also suggests that the rising incidence of student debt, resulting from the payment of tuition fees, might be deterring students from staying in academia. Although this mainly affects undergraduate intake, the report warns that graduate students saddled with debts may prove more reluctant to pursue an academic career.

In addition to stipend enhancements, the report recommends subsidizing housing and forgiving debt in some cases. Stipend top-ups, housing help and loan waivers all seem like sound ideas to recruit and retain young scientists — as long as principal investigators are made aware of them, and are encouraged to apply for the extra money.

Paul Smaglik, *Naturejobs* editor

Publisher: Ben Crowe
Editor: Paul Smaglik
Assistant Editor: Gene Russo

European Head Office, London
The Macmillan Building,
4 Crinan Street, London N1 9XW, UK
Tel: +44 (0) 20 7843 4961
Fax: +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

European Sales Manager:
Andy Douglas (4975)
e-mail: a.douglas@nature.com
Business Development Manager:
Amelie Pequignot (4974)
e-mail: a.pequignot@nature.com
Natureevents:
Claudia Paulsen Young (+44 (0) 20 7014 4015)
e-mail: c.paulsenyoung@nature.com
France/Switzerland/Belgium:
Muriel Lestringuez (4994)

UK/Ireland/Italy/RoW:
Loredana Milanese (4944)
Nils Moeller (4953)
Scandinavia/Spain/Portugal:
Evelina Rubio-Morgan (4973)
Germany/Austria/The Netherlands:
Reya Silao (4970)
Online Job Postings:
Matthew Ward (+44 (0) 20 7014 4059)

European Satellite Office
Germany: Patrick Phelan
Tel: +49 89 54 90 57 11
Fax: +49 89 54 90 57 20
e-mail: p.phelan@nature.com

Advertising Production Manager:
Stephen Russell
To send materials use London address above.
Tel: +44 (0) 20 7843 4816
Fax: +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

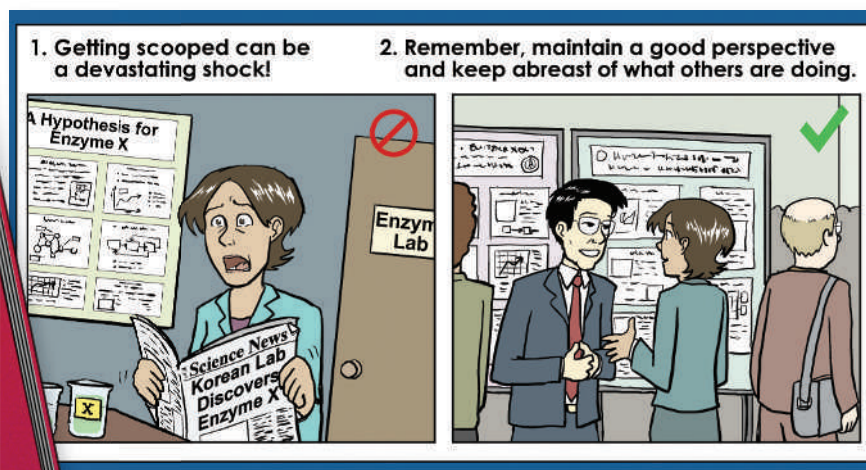
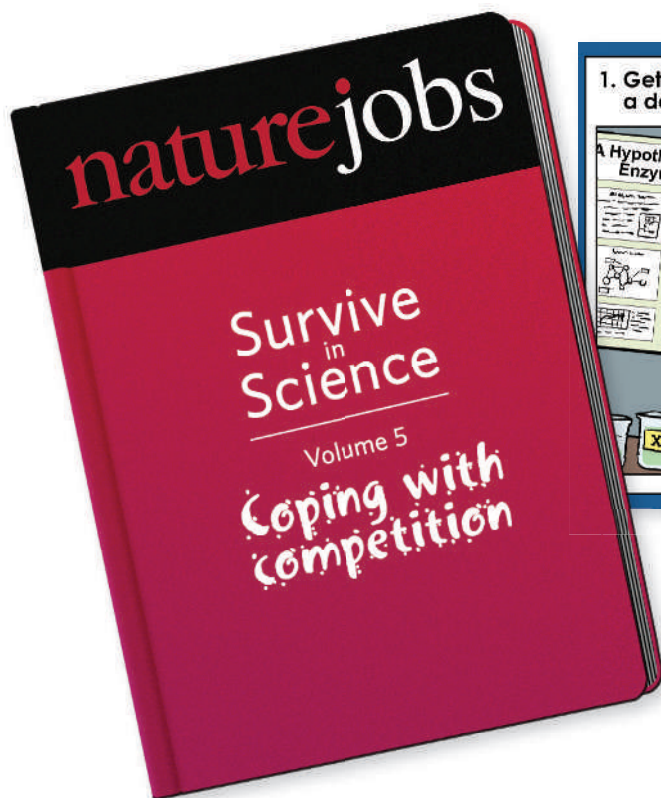
Naturejobs web development: Tom Hancock
Naturejobs online production:
Catherine Alexander

US Head Office, New York
75 Varick Street, 9th Floor,
New York, NY 10013-1917
Tel: +1 800 989 7718
Fax: +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

Japan Head Office, Tokyo
Chiyoda Building, 2-37 Ichigayatamachi,
Shinjuku-ku, Tokyo 162-0843
Tel: +81 3 3267 8751
Fax: +81 3 3267 8746

Asia-Pacific Sales Manager:
Ayako Watanabe
e-mail: a.watanabe@natureasia.com



Winning ways

After her thesis committee meeting, Dalia Halawani breathed a sigh of relief. Finally, after switching labs twice, she had a road map laid out for her thesis. Halawani, a PhD candidate at McGill University in Montreal, Canada, had spent the previous 18 months building the tools to investigate a key energy-transport enzyme. Two days later, a competing group published the answer in a paper that followed Halawani's plan exactly.

"I was absolutely shocked," says Halawani. "I thought that there was nothing left for me to address. I feared my competitor might be working on a follow-up paper."

Working on a project at the cutting edge can make a thesis or postdoctoral project worth the long hours and tedious experiments. But it also increases the odds that another group will find those answers first. In the growing competition for funding, high-profile publications and jobs, a shrinking violet is unlikely to succeed. "I was given the hottest project in the lab, but this is the disadvantage," says Halawani, of the project that had attracted such enthusiastic competition.

Preventive measures

"If you are working in a team that is very open with each other, you will benefit," says Didier Queloz, an astrophysicist at the University of Geneva in Switzerland. "At 25, you are selecting a subject that you know almost nothing about, but you can select the person you are going to work with and judge, 'Is this a nice guy?'" he suggests.

Ten years ago, during his postdoctoral fellowship at Geneva, Queloz worked with Michel Mayor on one of the hottest topics in astrophysics at the time, the hunt for

Science is cut-throat by nature, but how should young scientists handle working on competitive projects — or worse, getting scooped?

Kendall Powell investigates how to release the pressure valve.

the first extrasolar planet. He says that trusting Mayor, who kept him informed of developments in the field, was key to their successful discovery of planet 51 Pegasi b.

Queloz says that trust was never more evident than when he showed Mayor his results, which included evidence of a surprisingly large Jupiter-like planet with a tiny four-day orbit. His adviser asked him if he was confident in his calculations only once. When Queloz said he was sure, Mayor supported him and his data to ensure they published a strong paper.

"There will still be stress and pressure, but if you are in a good environment, that will translate into excitement and wanting to work the weekend," says Queloz.

Bruce Beutler, an immunologist at the Scripps Research Institute in La Jolla, California, agrees, noting that he often tells graduate students to hurry up on projects he knows are in a race with another group. He points out that working weekends means you are putting in 40% more effort than someone working only weekdays. "In a foot race, it never gets to be 40% apart — someone wins by just a few steps. People should really put everything in that they can," he says. Staying extremely focused is what led his group to be the first to show that a class of mammalian receptors sense bacterial invaders.

Molecular biologist Ming Tian at the University of Texas, Austin, says that joining a well-connected lab can help you stay ahead of the game. "You need to be informed of who is doing what," he says. If your adviser knows many people in the field, attends lots of meetings and has many collaborators, you will be in a better position to know what other groups are planning. In many disciplines, researchers hesitate to discuss unpublished results at conferences. Instead, much of this information comes from informal conversations at the pub, casual phone calls or the rumour mill.

If you know someone is way ahead of you, don't waste your time, Tian says. But he warns against getting too anxious about rumours: "If you have an idea you think is good, then just do your best to pursue it." Besides, he says, you never know when another group might get stuck with a technical problem or decide to switch directions. As a postdoctoral fellow, Tian worked on characterizing the DNA recombination events crucial to forming antibodies, competing with other groups that were doing the same thing. In fact, a competitor published during Tian's postdoc, but the paper was later retracted. Tian addressed the significant research question that the withdrawn paper left unanswered,

published his results, and became a much more competitive job candidate in 2003 as a result.

Students should pursue two types of project, suggests Martin Latterich, a cell biologist at McGill and Halawani's adviser. One should be high-profile, with a higher risk of getting scooped, and the other less ambitious but more likely to get published.

Having returned to academia from industry, Latterich also takes a lesson from the business world: "Create a niche that is truly unique and play to your strengths." For example, use a model system, assay or bioinformatics approach that no one else is using. Beutler has similar advice to avoid direct competition — use an approach such as random mutagenesis that will yield unexpected results instead of "deliberately looking for the most obvious things".

Scooped!

If the worst does happen, how does one recover? Gillian Wu, former dean of the science and engineering faculty at York University in Toronto, Canada, has a tiered strategy for her immunology group whenever a competing paper comes along. First, examine the paper with a fine-toothed comb and look for differences in your approaches, methods or results that your group could still capitalize on. If none arises, determine if your interpretations of the results differ enough to capitalize on those instead. Finally, if there are no differences overall, offer to collaborate with the group. "If you can't beat them, join them," she says. If this is not an option, move on to something new as quickly as possible.

Initially, Latterich and Halawani took a similar approach to getting scooped. First, they determined that the competing paper's results made sense, on the basis of their own preliminary data. Next, being a smaller lab, they decided they wouldn't compete head-to-head with their more established competitor by conducting the next most obvious experimental steps.

Halawani had already invested 18 months in the project, so abandoning it was not an option. She could have tried repeating her competitor's experiments to see if her results differed, but that was a gamble she did not want to take. Instead, she and Latterich decided to use techniques unique to their group to focus on areas where they could quickly publish.

"This whole incident forced me to be more creative," says Halawani. After getting over the shock of being scooped, she took a step back and wrote a review paper on the ATPase enzyme at the centre of her project. Doing a complete review of the literature, she learned which aspects of her project were most likely to be the focus of competing labs. "Now we are testing hypotheses and using models not used by people in this field," she says, "so even if someone publishes something similar, we will have quantified it differently."

Michael Alvarez, director of the Stanford School of Medicine Career Center in Palo Alto, California, says there's one more thing you should do for yourself if you get scooped: "Take inventory of the good things that have come out of the time you've spent on your project." This will remind you why you are doing this work in the first place, he says. The unpleasant reality of getting scooped does not lessen the skills and knowledge you gained.

It's crucial to keep competition in perspective, or the pressure will get you down. Tian says young scientists should remember that scientific rivalry is not personal

Dalia Halawani (right) rallied after burning her fingers on a hot project. **Michael Alvarez (below)** warns to watch out for signs of excessive pressure.



Martin Latterich: "Create a niche that is truly unique."



— it's just part of the game. Build up your contacts, don't burn them, adds Queloz. In any field, your competitors may turn out to be manuscript and grant reviewers, and even potential collaborators further down the road.

And being first is not synonymous with being successful. Although Beutler encourages his students and postdocs to hurry, he also warns them "not to shoot from the hip". Scientists who publish uncertain results simply to be first will quickly lose the respect of their colleagues, he warns.

In perspective

Latterich says your perspective on getting scooped changes as you progress. "The investigator sees a much bigger picture." Yes, some experiments were duplicated in another group, but now new angles will be pursued. At the same time, he acknowledges that getting scooped "devastated" Halawani, and that he had to switch into a more supportive adviser role, giving his normally independent student more encouragement and guidance.

Halawani says she copes with the pressure of competition by having lots of irons in the fire and investing herself 100% in finishing her doctorate. "If my protein expression fails, then my DNA cloning will work — that's how I manage."

Alvarez finds that a healthy amount of competitive pressure motivates people to perform better. It only becomes a problem when there is too little or too much. Pay attention to warning signs that the anxiety or stress may be reaching unhealthy levels, he advises.

"Watch out for both behavioural and mental state changes," he says. Are you eating too much or not enough? Trouble sleeping or getting out of bed? Hitting the bottle more often? Do you have feelings of despair or is your stomach in knots when you come into the lab? These are red flags that your stress levels are too high — talk to your adviser or another mentor about how you might adjust your project to cope better.

Luckily Halawani has a glass-half-full outlook, no doubt inspired by her adviser's attitude. "Don't be discouraged if you get scooped," Latterich says. "See it as a confirmation that what you are doing is important and interesting."

Kendall Powell is a freelance science writer based in Broomfield, Colorado.

My grandfather's river

A stream of memories.

Brenda Cooper

I begin to make the river. The river. His river. The one my grandfather took me down the year I turned 10, and again when I was 16 and 25.

It takes days to dig through web archives for his data, to find old versions of the two-dimensional geographic information software he used 20 years ago. Success allows me to form the data into three dimensions, to show the banks shift and the water fall away, to chart the demise of trees and animals.

It is not enough. Sitting straight in my chair, feet on the ground, back arched, stretching my wrists, I am tempted to give up and send a historybot to make a simple album of grandpa's speeches. Except his own words are no gift back to him, especially as they didn't work. He spent six years fighting to save the river, and then ten more wandering up and down it, studying his failure.

On our last trip, when I was 25 and he was 70, we sat in his red canoe in the middle of the river. A dead fish floated past us. "Why do you stay?" I asked.

"I need to save it."

I eyed the white underbelly of the dead fish but held my silence.

He looked away from me, his voice breaking. "I'm mapping it for you. I can't save the real river, but I can save the record of it." He pointed at a cloud of tiny cameras he'd set to follow us. Bright sun sparkled on them like diamonds.

I have recreated the river from that trip.

But I need the river of my youth, the one from our first trip.

I find a programbot that takes the old photographs from his first *Natural Geography* article and take two more days off work to scan, register and rectify the thin bright photos to his old 2D and make sure the cattails are exactly the right number of inches across, and that some bunch up close, hugging, and others wave above the water like brown flags.

I tell the pixellated water to rise up a little, watching carefully as depressions in

the banks fill into tiny spangled wetlands.

Olfactory databots yield pond water and cedar and frogs and mix them all up on command. I add throaty frog conversations, hoping this sensory stew will drive my little-girl memories forward. I collapse on the couch, the river surrounding me, washing me.

I cause stars of water spiders to scoot on bright drops of surface tension, and feed them digital mayflies. Virtual water laps at a finger I hold a few inches in front of my face. The water spiders glide and dance around it.



Finally, I slip into my child self and the memory of his voice is clear and strong, as if the river washed away the 40 years between then and now. "They never look like they're walking. Walking would be too slow for them."

I recall his hand on my shoulder as I gaze up into his intense blue-green eyes. He surveys the current, keeping us away from white water frothing over rocks. This man who is always gentle with me digs his fingers into my shoulder. The sun beats on his thinning blond hair as he lets go and makes such a sweeping gesture the canoe under us rocks alarmingly. "Penny. This is your heritage. We're stealing it from you. Memorize it, Penny. Memorize the water flowing always downstream, the clean, rounded rocks, the water spiders." Even the memory of his voice drives up details.

I add them one by one. Three turtles balancing on a floating log.

The ghostly feel of a warm wind.

A heron pretending to be a cattail.

The monitoring nano in my blood screams sleep at me and I can't override it any more without a doctor's chit.

It's okay. I'm done.

I collapse, sleeping for two days and a night, dreaming of turtles and herons and dragonflies.

On the morning of my grandfather's birthday I bring the river in my top pocket. The relentless sun beats down on the dry brown grass on his bit of lawn. He waits for me in an old wooden Adirondack chair, his eyes bright blue pools in a river of wrinkles from temple to temple. He smiles and

stands and holds me, his arms shaking a little. I suddenly hate it that he is 100 today.

Glancing down, I note that his nanomonitor is yellow again this morning. At least it isn't blinking in alarm.

Inside, a big white fan cools the kitchen, and there is no evidence he's eaten breakfast. He flips a switch and sits down, sighing in pleasure as the scent of brewing coffee puffs into the air, a history of mornings.

I stand behind him, kneading his shoulders, my throat tight. I slide the glasses out of my

pocket and slip them over his head.

His voice belongs to an old man. "What's this?"

My own voice shakes. "Look."

The glasses sense him and spring to life. Even though I can't see it, I know the river surrounds him. It runs over his ankles. Cattails grace the corner by the refrigerator. He grips his knee and a breath rushes from him. VR glasses are for an old man. I turn my retinas to virt. Reality greys to background. My senses catch up with the river programming just in time to be with him as the three turtles come into view in the empty doorframe.

He squeezes my palm hard.

I return the real world to my eyes.

A tear is falling down his cheek. ■

Brenda Cooper's stories have appeared in *Analog* and *Asimov's*, both solo and with collaborator Larry Niven. Her novel *The Silver Ship and the Sea* will be published by Tor in March 2007.

www.brenda-cooper.com

JACEY